SPATIO-TEMPORAL CHANGE-POINT DETECTION AND CONSTRAINED BAYESIAN OPTIMIZATION

A Thesis
Presented to
The Academic Faculty

by

Junzhuo Chen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2019

SPATIO-TEMPORAL CHANGE-POINT DETECTION AND CONSTRAINED BAYESIAN OPTIMIZATION

Approved by:

Professor Seong-Hee Kim, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yao Xie, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Mustafa M. Aral
Department of Civil Engineering
Bartin University

Professor Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Jianjun Shi
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: March 15, 2019

To my family, and all my friends.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Dr. Seong-Hee Kim and Dr. Yao Xie, for their support, encouragement and trust during these years. Their guidance helped me in all the time of research and writing of this thesis. From them, I learned how to become a researcher.

I would like to thank my thesis committee members: Dr. Mustafa M. Aral, Dr. Jianjun Shi and Dr. Kamran Paynabar for their kindness in evaluating my thesis and their valuable feedback. I would also like to thank Dr. Craig Tovey, Dr. Robert Foley, Dr. David Goldberg, Dr. Nagi Gebraeel, Dr. Alexander Shapiro and Dr. Yajun Mei for their excellent courses. In addition, I would like to thank Amanda Ford and Dr. Dawn Strickland for their dedicated services at ISyE.

I would like to thank all my fellow students and friends accompanying me during my study at Tech. The five years that I spent with them in Atlanta is so wonderful and will be such a precious memory in my life.

Last but not least, I would like to thank my parents for their unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis makes contributions to two research topics: spatio-temporal change-point detection and constrained Bayesian optimization. Spatio-temporal change-point detection is concerned with detecting statistical anomalies based on multiple data streams collected at different locations. In Chapter 2 and Chapter 3, we address two challenges in spatio-temporal change-point detection: (i) how to deal with data with high dimensionality, and (ii) how to capture spatial and temporal correlations. Bayesian optimization is a prevalent approach for optimization problems defined by expensive-to-evaluate black-box functions. In Chapter 4, we develop a practical algorithm for optimization problems with black-box objective function and constraints.

In Chapter 2, we study dimension reduction via spatial scanning. The majority of control charts using scan statistics for spatio-temporal change-point detection use full observation vectors. To deal with high dimensionality, most of the dimension reduction techniques are done as a post-processing step rather than in the data acquisition stage and thus the full sample covariance matrix is required. In a high dimensional application, (i) the sample covariance matrix tends to be ill-conditioned due to a limited number of samples; (ii) inversion of such a sample covariance matrix causes numerical issues; (iii) aggregating information from all variables may lead to high communication costs in sensor networks. We consider a set of reduced-dimension (RD) control charts which perform dimension reduction during data acquisition by spatial scanning and avoid the computational difficulties and possibly high communication costs. We characterize the performance difference between the RD and the full observation approaches, under several common spatial correlation models, in terms of average run lengths. Our results show that the RD approach has little performance loss under the correlation models considered in this chapter while enjoying all the implementation benefits. Our theoretical analysis is verified by extensive numerical studies including water quality monitoring.

In Chapter 3, we propose an efficient score statistic, called the $\mathsf{S}^3\mathsf{T}$ statistic, to detect the emergence of a spatially and temporally correlated signal from either fixed-sample or sequential data. The signal may cause a mean shift and/or a change in the covariance structure. The score statistic can capture both the spatial and temporal structures of the change, and hence, is particularly powerful in detecting weak signals. The score statistic is computationally efficient and statistically powerful. Our main theoretical contribution is analytical approximations of the false alarm rates of the detection procedures. Numerical experiments on simulated and real data, as well as a real case study of water quality monitoring, demonstrate the good performance of our procedure.

In Chapter 4, we study the problem of optimal sensor network design, which is formulated as a joint problem of constrained black-box function optimization and spatio-temporal change-point detection. We propose a practical algorithm called the Confidence-Set based Constrained Bayesian Optimization (CSCBO), which provides a flexible framework to handle noisy black-box function constraints and is easy to implement. We also extend the algorithm to tackle with a challenge that arises specifically in the sensor network design problem: we use the Wasserstein similarity metric to deal with high-dimensional binary decision variables. Finally, the $\mathsf{S}^3\mathsf{T}$ statistic proposed in Chapter 3 is combined with CSCBO to identify optimal sensor network designs that are robust to sensor measurement errors.

CHAPTER I

INTRODUCTION

The rapid development of sensor technology and communication network has enabled on-line monitoring of statistical anomalies based on high volume spatio-temporal data in a variety of industrial and service systems. The abrupt emergence of such an anomaly will change the distribution of the data and will usually cause destructive consequences, and hence quick detection is desired. Examples of spatio-temporal change-point detection include disease outbreak detection [30], water quality monitoring [1], and computer network intrusion detection [42]. While many methods have been established for classic problems, developing efficient detection procedure for spatio-temporal data involves new challenges. The first issue centers on the high-dimensionality of the data. For example, in a sensor network, the number of sensors deployed can be as large as thousands, which incurs high implementation cost if the data streams need to be processed centrally. Another issue is the complicated structure of the spatial and temporal correlations of the data. Usually the underlying process measured by the sensors is a spatio-temporal process where correlations exist in both space and time. Hence, the capacity of capturing correlation information is critical for efficient change-point detection procedures. In this thesis, we address these challenges through the following methods:

- We develop reduced-dimension methods via spatial scanning. Dimension reduction is achieved by breaking the monitored area into clusters and constructing local statistics for each cluster. The reduced-dimension methods with spatial scanning enjoy the following computational benefits: (i) it avoids estimating the entire sample covariance matrix which is difficult if we have a limited number of samples, (ii) it avoids the inversion of a large covariance matrix which is likely to cause numerical issues. On the other hand, the spatial scanning scheme also enables distributed implementation since data only needs to be processed locally instead of centrally. This will largely

reduce communication cost among sensors if the sensor network is deployed on a large area. Our theoretical and numerical studies show that the RD approach has little performance loss comparing to methods with full observation vectors and hence should be a preferable method in practice.

- We develop a new score statistic, called the $\mathsf{S}^3\mathsf{T}$ statistic, which captures both spatial and temporal correlation of a signal, and hence, is particularly powerful in detecting weak signals. The $\mathsf{S}^3\mathsf{T}$ statistic also avoids inversion of the spatial-temporal covariance matrix and is hence computational efficient comparing to the maximum likelihood ratio statistic. We also develop theoretical approximations of the false alarm rates of the proposed detection procedures. Performance of the new methods is demonstrated via simulated data and real data.

In this thesis, we also study the problem of optimal sensor network design, which consists of two components: selection of sensor locations and design of detection statistics. The sensor network we study is used for water quality monitoring, and hence, the design of such a network should answer not only where the sensors should be placed in space but also how the data collected should be processed and analyzed. The objective is to achieve minimal detection delay when a contamination event occurs, and meanwhile, the sensor network is subject to the constraints on the probability of detection and the false alarm rate. We formulate the problem as a joint problem of constrained black-box function optimization and spatio-temporal change-point detection. Black-box function optimization refers to the optimization problems defined by functions that do not have analytical forms and usually can only be evaluated via computer simulations. Bayesian optimization (BO) is a prevalent method for unconstrained black-box function optimization problems. In Chapter 4, we develop a practical algorithm called the Confidence-Set based Constrained Bayesian Optimization (CSCBO), which can be used to solve problems with black-box function constraints. We also use the Wasserstein similarity metric to tackle with high-dimensional binary decision variables, which is a challenge that arises specifically in the sensor network design problem. Finally, we combine the $\mathsf{S}^3\mathsf{T}$ statistic with CSCBO to identify optimal

sensor network designs that are robust to sensor measurement errors.

1.1 Literature Review

1.1.1 Spatio-temporal change-point detection

Multivariate control charts are commonly used for spatio-temporal change-point detection. From a stream of observation vectors, a control chart computes a sequence of monitoring statistics and detects a change whenever the monitoring statistic goes beyond certain control limits. The control limits are pre-specified according to the requirement for the in-control average run length (known as the ARL_0), which captures the false-alarm-rate of a control chart. Another related performance metric is the out-of-control average run length (known as the ARL_1), which represents the expected number of samples needed to raise an alarm. Our goal is to detect the change quickly, namely, to achieve a short detection delay or ARL_1 for a given targeted ARL_0.

Classical multivariate control charts mainly consider low-dimensional problems with a few number of data streams. Commonly used multivariate control charts include the T^2 chart by [21], the multivariate exponentially weighted moving average (MEWMA) chart by [34], and the multivariate cumulative sum (MCUSUM) chart. The CUSUM chart [38] is widely adopted due to its good property in detecting small shifts and its efficient recursive implementation that facilitates online monitoring. In the multivariate setting, one may construct MCUSUM charts by directly using vector observations; such methods can be largely classified as the log likelihood ratio (LR) based and the Hotelling's T^2 statistic based methods. Among them, [19] uses a log LR statistic assuming the vector observations are i.i.d. multivariate normal. [11] computes a T^2 statistic for each vector observation and then forms a CUSUM chart based on a sequence of T^2 statistics. A key difference between the LR based and the T^2 based MCUSUM statistics is that the former assumes known shift direction vectors, while the latter does not make such an assumption.

Compared with the settings of classical multivariate control charts, spatio-temporal change-point detection tends to be much higher dimensional. One important instance of spatial-temporal change-point detection is disease outbreak detection, where a decision

maker collects health measurements such as disease counts or mortality rate from a number of adjacent regions to monitor a potential public health hazard (see [68] for a review). For example, [61] consider a problem of simultaneously monitoring 200,000 indicators of excess mortality in the UK health system. In this setting, various MCUSUM charts have been developed that describes the spatial disease pattern in the entire monitored area. See, for example, [47] and [64]. Since a disease outbreak usually happens in a group of neighboring regions, [60] incorporates spatial scanning and defines a spatial cluster as a group of local regions lying within a circle of certain radii, which is an extended formulation of [48]. However, spatial correlation is not considered in [60]. Adopting a similar definition of spatial clusters in [60], [22] propose an MCUSUM chart, which constructs an LR statistic for each spatial cluster under the normal assumption and scans through all possible clusters to detect a possible outbreak. [30] revise this method by using an analytical formula developed by [25] to approximate the control limits. Then, [29] extend the previous work to more general distributions. Besides public health monitoring, [71] propose a generalized-likelihood-ratio-test control chart for monitoring a product surface data with the assumption that the surface data follow the multivariate normal distribution with an identity covariance matrix to detect a mean shift of a certain pattern. Another instance of high-dimensional spatial temporal change-point detection occurs in astronomical imaging, where high-resolution video streams are monitored for solar flare detection, as studied in [72] and [33]. In this problem, each observation is a 67,744 dimensional vector consisting of image pixels and the goal is to detect an emergence of a sparse signal. Sequential detection of a sparse change is an active area of research, see, e.g., the work by [74] and [32]. However, these work usually assumes independent data streams without considering spatial correlation. See, for example, [12] for the importance of capturing spatial correlation for effective monitoring.

To tackle the difficulties caused by high-dimensionality, a viable solution is to perform a dimensionality reduction technique. The existing dimension reduction techniques include the principal component analysis (PCA) such as [35]; random linear projections such as [51], [3] and [58]; and wavelet transform methods such as [28] and [70]. In addition, [18] propose a projection based method combined with a T^2 control chart, and [6] and [77] propose

a Bayesian hierarchical approach to dimension-reduced spatio-temporal modeling for the task of predicting a high-dimensional response from a high-dimensional predictor. These methods still require full observation vectors because dimension-reduction is performed as a *post-processing* step rather than *in the data acquisition stage*. In other words, while monitoring a process, full observation vectors need to be collected and thus high communication cost is still incurred in addition to a possibly ill-conditioned full sample covariance matrix.

1.1.2 Optimal sensor network degisn

The optimal selection of sensor locations for water quality monitoring network has been studied by many researchers. [63] provides a comprehensive review of past approaches. Among the recent works, [66] and [67] formulate the problem as a optimization problem with two objectives: minimizing the detection delay and maximizing the probability of detection. A genetic algorithm (GA) is used to solve the optimization problem. [40] and [41] formulate the problem as a constrained optimization problem to minimize the detection delay with a constraint on the probability of detection and adopt a combined procedure of the nested partition (NP) [52] and the penalty function with memory [39]. We refer to the combined procedure as NP + PFM, which is proved to converge almost surely to the true optimal feasible solution [39].

Optimal sensor placement is essentially a black-box function optimization problem as the objective function (detection delay) and the constraint (probability of detection) do not have analytical forms and can only be evaluated via simulations. Bayesian optimization (BO) [36] is a prevalent method for black-box function optimization problems. A BO algorithm typically models a function by a Gaussian process (GP) [45] and guide the search based on an acquisition function. [59] provides an overview on how BO algorithms are applied in various tasks. Recently, some researchers extend BO algorithms to optimization problems with black-box function constraints. These works include [14], [2], [20], [27] and [31]. Typically, applications of BO algorithms have continuous decision variable with relatively low dimensionality. However, in the problem of sensor network design, the decision variable is the set of sensor locations and hence is a vector of categorical variable. Thus,

a reasonable similarity metric over sets of locations is critical to apply BO algorithms for sensor network design. [15] propose to use the earth mover's distance as the similarity metric and apply a unconstrained BO algorithm to find the optimal locations for weather sensors in the United Kingdom.

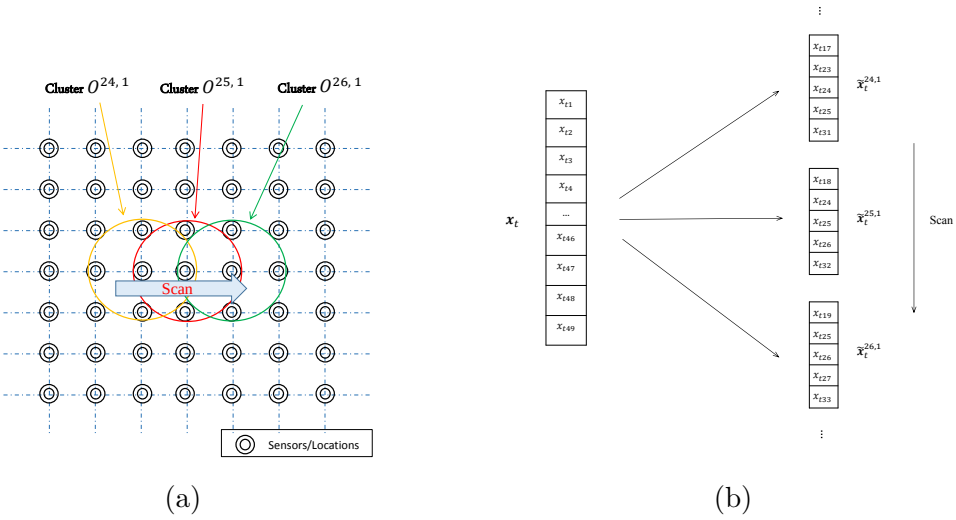
CHAPTER II

TO REDUCE OR NOT TO REDUCE: A STUDY ON SPATIO-TEMPORAL CHANGE-POINT DETECTION

In spatio-temporal change-point detection, one monitors an area using data streams measured at different locations and aims to detect any changes as soon as they occur. Typically, the data streams are observations of a particular quality index, which are sequentially collected by either physical sensors or professionals from different locations in the monitored area. In most cases, the quality index at each location can be modeled as either a discrete (e.g., network intrusion counts, or mortality) or a continuous (e.g., disease incidence rate, or contaminant concentration) random variable.

For high-dimensional applications, all control charts reviewed in Section 1.1.1 use full observation vectors when constructing monitoring statistics. Even for charts with the spatial scanning approach ([22]; [30] and [29]), a full-size covariance matrix is involved in computing the detection statistic. In the high-dimensional setting in the presence of spatial correlation, several difficulties exist in applying these control charts that require a full-size covariance matrix to real scenarios: (i) the sample covariance matrix tends to be ill-conditioned due to the number of samples being relatively small compared to the dimension of the covariance matrix in the streaming setting; (ii) the ill-conditioned sample covariance matrix causes numerical issues due to the matrix inversion involved in computing the statistics; and (iii) communication cost can be high for distributed sensor networks, since using full observation vectors means that all sensors need to exchange information with each other [17].

Another way to tackle the high-dimensionality problem is to perform *reduced-dimension* spatial scanning. One first breaks the entire monitoring area into overlapping local clusters of certain radii and only uses a subset of sensors or locations within the clusters, as in Figure 1. Then, one constructs a control-chart for each local cluster and detects a change whenever any of the local clusters fires an alarm. By doing so, each control chart only



(a) (b)

Figure 1: (a) A monitored area with $p = 7 \times 7$ locations or sensors and illustration of the spatial scanning using a circular shaped region; (b) Mapping of a full-dimensional observation vector into reduced-dimensional vectors corresponding to scanning regions.

monitors a small number of data streams that fall within the scanning cluster. In this chapter, we consider the reduced-dimensional spatial scanning with the spatial correlation *within the local cluster* only, rather than the full spatial correlation as in the earlier works such as [22]. Hence, our method, due to its spatial scanning nature, never needs to acquire full-dimensional observations and dimension reduction is performed *during data acquisition*. Our method is suitable for distributed processing required by sensor networks. A recent related work by [75] based on linear projection may also be used for spatial scanning; however, spatial correlation is not considered in that work. Moreover, due to the recursive nature of the CUSUM statistics, our method is suitable for *in situ* processing, which means that raw data are not needed to be stored, and this is again preferable in sensor networks.

Although the idea of the reduced-dimension spatial scanning is not new and, in fact, is often used in practice, one important question has never been unanswered in the literature: *How much do we lose by using reduced-dimension observations in the presence of spatial correlation?* We provide a precise answer to the amount of loss, by characterizing the difference of reduced-dimension charts and full-dimensional charts in term of their ARL_1 under a fixed ARL_0. Our analysis shows that the RD approach usually spends $0 \sim 20\%$ more observations in ARL_1 than the full observation approach for a reasonable range of spatial

correlation among neighboring regions. Even when we lose more than 20%, the absolute differences in ARL_1 are small as one to five observations in most cases. In addition, we show that it is even possible that the RD approach may perform better than the full observation approach (i) when T^2 based charts for unknown shift directions are used or (ii) when some groups are completely independent of other groups as in a water quality monitoring example presented in Section 2.5. This is a blessing since an anomaly usually affects a local region and spatial correlation tends to decay with a distance. Thus, restricting to local sensors when forming a monitoring statistic should not significantly degrade the detection performance.

2.1 Background

In this section, we define notation and our problem. Then a few spatial scanning control charts are presented as representative charts that take the full observation approach.

2.1.1 Notation and Problem

In spatio-temporal change-point detection, observations (e.g., a quality index) are sequentially collected from different locations in the monitored area. Using the sequential observations, one desires to detect a possible change or anomaly in the monitored area as soon as possible.

Suppose there are p locations (sensors). For simplicity, we assume that the monitored area is rectangular and the locations (sensors) sit at a lattice of $p = MN$ points. This rectangular assumption on the shape of the monitored area can be relaxed, as we show in Section 2.5. Let $q_c = (m_c, n_c)$, where $m_c = 1, 2, \ldots, M$ and $n_c = 1, 2, \ldots, N$ denote the two-dimensional spatial coordinate of the location indexed by c, and $c = (m_c - 1)N + n_c$. Let $P = \{1, 2, \ldots, p\}$ be the set of monitoring locations. At time t, the observation is a p-dimensional vector $\boldsymbol{x}_t = [x_{t1}, x_{t2}, \ldots, x_{tp}]'$. Assume that different observation vectors are temporally independent but spatially correlated with covariance matrix $\boldsymbol{\Sigma}$. Further assume that the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}_t$ is known or can be estimated from data. The change only affects the mean and the covariance matrix remains the same. Under the hypothesis of no change, the observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ are i.i.d. normally distributed with a mean vector $\boldsymbol{\mu}_0$ and a covariance matrix $\boldsymbol{\Sigma}$. Alternatively, there exists a change-point κ in

time, which represents an anomaly, and a subset of neighboring locations are *affected* by the change-point. For the locations affected by the change, the means of their observations are shifted, while observations from the unaffected locations keep the same distribution. This corresponds to a shift in the mean vector from $\boldsymbol{\mu}_0$ to some other vector $\boldsymbol{\mu}_1$. Without loss of generality, assume that the observation vectors have been standardized so that $\boldsymbol{\mu}_0 = \mathbf{0}$ and $[\boldsymbol{\Sigma}]_{i,i} = 1, \forall i = 1, \ldots, p$, where $[\cdot]_{i,j}$ denotes the (i,j)th element of a matrix.

Due to spatial correlation, an anomaly often affects a cluster of neighboring locations. We assume that the shift cluster is circular shaped to facilitate the notation (but the cluster does not need to be circular shaped, which will be demonstrated in Section 2.5). A cluster is a set of locations $O^{c,r} = \{j | \|q_j - q_c\| \leq r,\ j \in P\}$, where c is the center of the cluster, r is the radius of the cluster, where $\|\cdot\|$ denotes the ℓ_2-norm of a vector. In our setting (sensors are placed over a grid), the radius r is usually chosen from a discrete set of values. Let $R \subseteq \{1,\ \sqrt{2},\ 2,\ 2\sqrt{2}, \cdots\}$ be the set of possible values of r. Define an p-dimensional vector $[\boldsymbol{\mu}_{c,r}]_j = \delta_j$ for all $j \in O^{c,r}$ and 0, otherwise. Here, $[\cdot]_j$ denotes the jth element of a vector, and δ_j denotes the shift magnitude of the jth location. Hence, if an anomaly affects the cluster $O^{c,r}$, then $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{c,r}$.

Our goal is to detect anomalies that affect a cluster $O^{c,r}$, $c \in P$ and $r \in R$ by testing whether the mean of the observations has shifted from a nominal vector $\boldsymbol{\mu}_0$ to a different vector $\boldsymbol{\mu}_1$.

2.1.2 LR-F-MCUSUM chart

[22] consider a spatial scanning control chart based on LR statistics. Their method is based on full observation vectors and needs a full covariance matrix. Spatial scanning is achieved by zeroing out the part of the mean vector that falls out of the scanning region. Below, we review their method, referred to as the LR-F-MCUSUM chart hereafter. For a hypothetical shift cluster with center c and radius r, the sequence of monitoring statistics is given by,

$$S_t^{c,r} = \max_{1 \leq \tau \leq t} \sum_{i=\tau}^{t} \ell_i^{c,r}, \qquad t = 1, 2, \ldots, \tag{1}$$

where $\ell_i^{c,r}$ is the log LR statistic when the post-change mean vector is $\boldsymbol{\mu}_{c,r}$:

$$\ell_i^{c,r} = \boldsymbol{\mu}_{c,r}^{'} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}_i - \frac{\boldsymbol{\mu}_{c,r}}{2} \right), \quad i = 1, 2, \ldots, t. \tag{2}$$

When constructing the monitoring statistic in (1), we need to search over the unknown change-point time by maximizing with respect to τ, the variable that represents a putative change-point location. Note that the statistic for each cluster $S_t^{c,r}$ in (1) can be computed recursively

$$S_t^{c,r} = \max\{0, S_{t-1}^{c,r} + \ell_t^{c,r}\}, \quad t = 1, 2, \ldots, \quad \text{and} \quad S_0^{c,r} = 0. \tag{3}$$

If the true sfhit center and radius (c, r) are known, we may perform the change-point detection when $S_t^{c,r}$ exceeds a pre-specified control limit. In practice, usually neither the shift center c nor the radius r is known a priori. In this case, one has to scan over all possible values of c and r, calculate the corresponding MCUSUM statistic, and form a global detection statistic by taking the maximum: $S_t^{**} = \max_{c \in P, r \in R} S_t^{c,r}$ for $t = 1, 2, \ldots$. A change is detected whenever S_t^{**} exceeds a pre-specified control-limit $h_\ell^{**} > 0$, which is specified according to the requirement for the ARL_0.

2.1.3 T^2-F-MCUSUM chart

When there is no prior information on the magnitude or direction of the mean shift, a T^2-statistic based MCUSUM chart is more appropriate. To have a reasonable comparison with reduced dimension charts, we introduce a T^2-F-MCUSUM chart based on full dimensional observation vectors. It performs spatial scanning while using the full covariance matrix as analogous to [22]. Given a cluster $O^{c,r}$, we modify the full observation vector $\boldsymbol{x}_t$ by replacing all elements that are not in the cluster with zeros $[\boldsymbol{x}_t^{c,r}]_j = [\boldsymbol{x}_t]_j$ for $j \in O^{c,r}$ and 0, otherwise. For each modified observation $\boldsymbol{x}_t^{c,r}$, compute a T^2 statistic

$$a_t^{c,r} = \boldsymbol{x}_t^{c,r\,'} \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t^{c,r} - \mu_T - k\sigma_T, \tag{4}$$

where k is a positive real-valued constant, $\mu_T = \text{E}[\boldsymbol{x}_t^{c,r\,'} \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t^{c,r}]$ and $\sigma_T^2 = \text{Var}[\boldsymbol{x}_t^{c,r\,'} \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t^{c,r}]$, which are the in-control mean and standard deviation of the T^2 statistic, respectively. The calculations for μ_T and σ_T^2 are discussed in Section 4. The monitoring statistic $T_t^{c,r}$ is

computed recursively over time

$$T_t^{c,r} = \max\{0, T_{t-1}^{c,r} + a_t^{c,r}\}, \quad t = 1, 2, \ldots, \quad \text{and} \quad T_0^{c,r} = 0. \tag{5}$$

With unknown shift center and radius, we again search over all possible clusters and sizes to form the global detection statistic $T_t^{**} = \max_{c \in P, r \in R} T_t^{c,r}$ for $t = 1, 2, \ldots$ and detection is performed by comparing T_t^{**} with a control limit h_a^{**}.

2.2 RD-MCUSUM Charts

In this section, we present the reduced-dimension approach. For each scan cluster, a control chart is constructed for reduced-dimension observation vectors while only considering local covariance. We develop two versions of MCUSUM charts, based on the LR statistic and the T^2 statistic, respectively.

2.2.1 LR-RD-MCUSUM chart

We start by considering an LR based chart, which is referred as the LR-RD-MCUSUM chart hereafter. For each scan cluster $O^{c,r}$, we truncate the original data vector $\boldsymbol{x}_t$ into a lower dimensional vector $\tilde{\boldsymbol{x}}_t^{c,r}$, where $[\boldsymbol{x}_t]_j$ is positioned in $\tilde{\boldsymbol{x}}_t^{c,r}$ if $j \in O^{c,r}$, and is eliminated otherwise, as illustrated in Figure 1(b). The monitoring statistic for that particular cluster is computed over vectors of dimension $|O^{c,r}|$, where $|\cdot|$ denotes the cardinality of a set. At each time, the LR statistic for $O^{c,r}$ is computed

$$\tilde{\ell}_i^{c,r} = \tilde{\boldsymbol{\mu}}_{c,r}^{'} \boldsymbol{\Sigma}_{c,r}^{-1} (\tilde{\boldsymbol{x}}_i^{c,r} - \frac{\tilde{\boldsymbol{\mu}}_{c,r}}{2}), \quad i = 1, 2, \ldots, t. \tag{6}$$

Here $\tilde{\boldsymbol{\mu}}_{c,r}$ and $\boldsymbol{\Sigma}_{c,r}$ are the sub-vector and sub-matrix of $\boldsymbol{\mu}_{c,r}$ and $\boldsymbol{\Sigma}$, respectively. Then, based on (6), the detection statistic $\tilde{S}_t^{c,r}$ is computed recursively for each cluster similar to (3). Finally, a global monitoring statistic is formed by taking the maximum over all clusters and sizes

$$\tilde{S}_t^{**} = \max_{c \in P, r \in R} \tilde{S}_t^{c,r}, \quad t = 1, 2, \ldots. \tag{7}$$

An alarm is signaled whenever $\tilde{S}_t^{**}$ exceeds a pre-specified control limit $\tilde{h}_\ell^{**}$.

Table 1: Summary of charts with the full and RD approaches.

	F-MCUSUM	RD-MCUSUM
LR Based	$S_t^{**} = \max_{c,r} S_t^{c,r} \geq h_\ell^{**}$	$\tilde{S}_t^{**} = \max_{c,r} \tilde{S}_t^{c,r} \geq \tilde{h}_\ell^{**}$
T^2 Based	$T_t^{**} = \max_{c,r} T_t^{c,r} \geq h_a^{**}$	$\tilde{T}_t^{**} = \max_{c,r} \tilde{T}_t^{c,r} \geq \tilde{h}_a^{**}$

2.2.2 T^2-RD-MCUSUM chart

Finally, similar to the reduced-dimension LR chart above, we construct the reduced-dimension T^2 statistic chart, which is referred to as the T^2-RD-MCUSUM chart hereafter. Given a cluster, at each time, we compute a T^2 statistic,

$$\tilde{a}_t^{c,r} = \tilde{\boldsymbol{x}}_t^{c,r'} \boldsymbol{\Sigma}_{c,r}^{-1} \tilde{\boldsymbol{x}}_t^{c,r} - \tilde{p} - k\sqrt{2\tilde{p}}, \quad i = 1, 2, \ldots, t, \tag{8}$$

where $\tilde{p} = |O^{c,r}|$. Note that when the process is in-control, the mean and variance of T^2 statistic is given by

$$\mathrm{E}[\tilde{\boldsymbol{x}}_t^{c,r\ \prime} \boldsymbol{\Sigma}_{c,r}^{-1} \tilde{\boldsymbol{x}}_t^{c,r}] = \tilde{p}, \quad \text{and} \quad \mathrm{Var}[\tilde{\boldsymbol{x}}_t^{c,r\ \prime} \boldsymbol{\Sigma}_{c,r}^{-1} \tilde{\boldsymbol{x}}_t^{c,r}] = 2\tilde{p}.$$

Based on $\tilde{a}_t^{c,r}$, the monitoring statistic for each cluster is computed recursively over time similar to (5). The global detection statistic is formed by maximizing over all clusters: $\tilde{T}_t^{**} = \max_{c \in P, r \in R} \tilde{T}_t^{c,r}$ for $t = 1, 2, \cdots$. An alarm is signaled when $\tilde{T}_t^{**}$ exceeds a control limit $\tilde{h}_a^{**}$.

Hereafter, the charts based on full observation vectors are referred to as the F-MCUSUM charts, including LR-F-MCUSUM and T^2-F-MCUSUM charts, and the reduced dimension charts are referred to as the RD-MCUSUM charts, including LR-RD-MCUSUM and T^2-RD-MCUSUM. Table 1 summarizes our methods and terminology.

2.3 Theoretical Analysis for Effects of Spatial Correlation

In this section, we compare analytically the performance of the F-MCUSUM charts with that of the RD-MCUSUM charts. We use ARL_1 given a target ARL_0 as our performance metric. To make a fair comparison among various control charts, we calibrate the control limits so that their actual ARL_0 are equal to the target values.

Both the F-MCUSUM and RD-MCUSUM charts use scan statistics to search for the true shift cluster from the set of all possible shift clusters $\{O^{c,r}|c \in P,\ r \in R\}$ at each time. In practice, the center and the radius of an anomaly are unknown so the entire monitored area needs to be scanned. In this section, we perform a theoretical analysis of the F-MCUSUM and RD-MCUSUM charts in a simplified setting, i.e., when the actual shift cluster is known. This simplified situation provides some insights into understanding the impact of dimensionality reduction on the performance of an MCUSUM chart with scanning.

Suppose a shift affects a cluster with center c and radius r, respectively, i.e., $O^{c,r}$ is the actual shift cluster. Suppose the cluster $O^{c,r}$ contains $\tilde{p}$ locations. Without loss of generality, assume the affected locations correspond to the first $\tilde{p}$ entries in the observation vectors, for instance, through reindexing. Hence, the post-change mean vector is $\boldsymbol{\mu}_1 = [\mu_1, \ldots, \mu_{\tilde{p}}, 0, \ldots, 0]' = [\tilde{\boldsymbol{\mu}}'_{c,r}, 0, \ldots, 0]'$, where $\tilde{\boldsymbol{\mu}}'_{c,r} = [\mu_1, \ldots, \mu_{\tilde{p}}]' \neq \mathbf{0}$. We have $\boldsymbol{x}_t = [\tilde{\boldsymbol{x}}_t^{c,r'}, 0, \ldots, 0]' + [0, \ldots, 0, \hat{\boldsymbol{x}}_t^{c,r'}]' = \boldsymbol{x}_t^{c,r} + [0, \ldots, 0, \hat{\boldsymbol{x}}_t^{c,r'}]'$, where $\tilde{\boldsymbol{x}}_t^{c,r} = [x_{t1}, \ldots, x_{t\tilde{p}}]'$ and $\hat{\boldsymbol{x}}_t^{c,r'} = [x_{t(\tilde{p}+1)}, \ldots, x_{tp}]'$. Furthermore, partition the $p \times p$ dimensional covariance matrix for the full observation vectors accordingly $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, where $\boldsymbol{\Sigma}_{11} \in R_{\tilde{p}\times\tilde{p}}$, $\boldsymbol{\Sigma}_{12} \in R_{\tilde{p}\times(p-\tilde{p})}$, $\boldsymbol{\Sigma}_{21} \in R_{(p-\tilde{p})\times\tilde{p}}$ and $\boldsymbol{\Sigma}_{22} \in R_{(p-\tilde{p})\times(p-\tilde{p})}$. Using the Schur complement [78] of $\boldsymbol{\Sigma}_{11}$, we write the inverse of $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{A^*}^{-1} & \boldsymbol{\Sigma}_B^{-1} \\ \boldsymbol{\Sigma}_C^{-1} & \boldsymbol{\Sigma}_D^{-1} \end{bmatrix}, \qquad \text{where}$$

$$\begin{aligned} &\boldsymbol{\Sigma}_{A^*}^{-1} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}; \quad \boldsymbol{\Sigma}_B^{-1} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}; \\ &\boldsymbol{\Sigma}_C^{-1} = -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}; \quad \text{and} \quad \boldsymbol{\Sigma}_D^{-1} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}. \end{aligned} \tag{9}$$

2.3.1 Relations between statistics in full and reduced-dimension charts

As a basis for the subsequent analysis, we derive relations for the LR and T^2 statistics used in the F- and the RD-MCUSUM charts. Recall that the likelihood ratio and the T^2 statistic based on full observations are $\ell_t^{c,r}$ and $a_t^{c,r}$, respectively and that their counterparts based

on reduced-dimension vectors are $\tilde{\ell}_t^{c,r}$ and $\tilde{a}_t^{c,r}$, respectively. Using equations (2), (6) and (9), the original LR statistic (2) can be written as

$$\begin{aligned}\ell_t^{c,r} &= \tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{x}}_t^{c,r} - 0.5\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r} + \tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{x}}_t^{c,r} + \tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{B}^{-1}\hat{\boldsymbol{x}}_t^{c,r} - 0.5\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{\mu}}_{c,r} \\ &= \tilde{\ell}_t^{c,r} + \tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{x}}_t^{c,r} + \tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{B}^{-1}\hat{\boldsymbol{x}}_t^{c,r} - 0.5\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}.\end{aligned} \tag{10}$$

From (10), we see that dimension reduction for LR based chart is equivalent to removing the last three terms on the right hand side of (10) from the full LR statistic. It is worth mentioning that $\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{B}^{-1}\hat{\boldsymbol{x}}_t^{c,r}$ contains only noise information since no mean shift occurs in $\hat{\boldsymbol{x}}_t^{c,r}$.

Similarly, we derive a relation for the T^2 statistics in the F- and the RD-MCUSUM charts. By equation (9) we obtain the in-control mean and variance of the full version T^2 statistic as follows:

$$\begin{aligned}\mathrm{E}[\boldsymbol{x}_t^{c,r\,'}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t^{c,r}] &= \tilde{p} + \mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} \quad \text{and} \\ \mathrm{Var}[\boldsymbol{x}_t^{c,r\,'}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t^{c,r}] &= 2\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 4\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 2\tilde{p}.\end{aligned}$$

Hence, the statistics $a_t^{c,r}$ defined in (4) and $\tilde{a}_t^{c,r}$ defined in (8) are related to each other via

$$\begin{aligned}a_t^{c,r} &= \tilde{\boldsymbol{x}}_t^{c,r'}(\boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{A^*}^{-1})\tilde{\boldsymbol{x}}_t^{c,r} - (\tilde{p} + \mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}) - k[2\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 4\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 2\tilde{p}]^{1/2} \\ &= \tilde{a}_t^{c,r} + \tilde{\boldsymbol{x}}_t^{c,r'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{x}}_t^{c,r} - \mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} - k([2\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 4\mathrm{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\} + 2\tilde{p}]^{1/2} - (2\tilde{p})^{1/2}).\end{aligned}$$

2.3.2 Performance metric: $\mathbf{ARL_1}$ measure

We aim to compare the detection performance of the F- and the RD-MCUSUM charts in terms of their ARL_1 for a fixed ARL_0. To do so, we will define a performance metric called the *ARL_1 measure*. For a fixed ARL_0, a smaller ARL_1 measure implies a smaller ARL_1.

[25] derive a formula to approximate ARL for both in-control and out-of-control processes of a single CUSUM chart, which can be used to approximate ARL_1 for a fixed target value, ARL_0. The formula is given by

$$\mathrm{ARL} \approx \begin{cases} \dfrac{\Omega^2}{2d^2}\left\{\exp\left[-\dfrac{2d(H+1.166\Omega)}{\Omega^2}\right] - 1 + \dfrac{2d(H+1.166\Omega)}{\Omega^2}\right\}, & \text{if } d \neq 0; \\ \left(\dfrac{H+1.166\Omega}{\Omega}\right)^2, & \text{if } d = 0, \end{cases} \tag{11}$$

Table 2: Drift and variance parameters of the LR based charts.

	LR-F-MCUSUM	LR-RD-MCUSUM
d_0	$-\frac{1}{2}\boldsymbol{\mu}_{c,r}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{c,r}$	$-\frac{1}{2}\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}$
$d_{c,r}$	$\frac{1}{2}\boldsymbol{\mu}_{c,r}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{c,r}$	$\frac{1}{2}\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}$
$\Omega_0^2(\Omega_{c,r}^2)$	$\boldsymbol{\mu}_{c,r}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{c,r}$	$\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}$

where d is the drift parameter, Ω^2 is the variance parameter, and H is the control limit.

In our settings, specifically, if observations are temporally independent, for the likelihood ratio statistics based chart, $d = \mathrm{E}[\ell_t^{c,r}]$, $\Omega^2 = \mathrm{Var}[\ell_t^{c,r}]$ for the LR-F-MCUSUM chart, and $d = \mathrm{E}[\tilde{\ell}_t^{c,r}]$, $\Omega^2 = \mathrm{Var}[\tilde{\ell}_t^{c,r}]$ for the LR-RD-MCUSUM chart; in the T^2 based charts, $d = \mathrm{E}[a_t^{c,r}]$, $\Omega^2 = \mathrm{Var}[a_t^{c,r}]$ for the T^2-F-MCUSUM chart and $d = \mathrm{E}[\tilde{a}_t^{c,r}]$, $\Omega^2 = \mathrm{Var}[\tilde{a}_t^{c,r}]$ for the T^2-RD-MCUSUM chart. In the following, we denote the in-control drift and variance parameters as d_0 and Ω_0^2, respectively. The out-of-control parameters are defined similarly. If the shift center is c and radius is r, we denote the out-of-control drift and variance as $d_{c,r}$ and $\Omega_{c,r}^2$, respectively. Tables 2 and 3 summarize these parameters.

The in-control drift d_0 is always negative but the out-of-control drift $d_{c,r}$ can be either negative (if the shift magnitude is too small) or positive. In this section, we assume that the shift magnitude is large enough so that $d_{c,r}$ is positive.

Consider a special function called the Lambert W function. Let $W_0(\cdot)$ be the principal branch of the Lambert W function [10]. For fixed target ARL_0, $d_0 < 0$ and $d_{c,r} > 0$, using (11) and the notations defined above, we derive an approximation for ARL_1 as follows:

$$\mathrm{ARL}_1 \approx -\frac{\Omega_0^2}{2d_0 d_{c,r}}\epsilon_{\eta_0} - 1.166\frac{1}{d_{c,r}}(\Omega_0 - \Omega_{c,r}), \tag{12}$$

where

$$\epsilon_{\eta_0} = -W_{-1}(-e^{-\eta_0}) - \eta_0, \quad \text{and} \quad \eta_0 = \frac{2d_0^2}{\Omega_0^2}\mathrm{ARL}_0 + 1.$$

The derivation is presented in Appendix A.1. It is noteworthy that the ϵ_{η_0} function tends to be very flat. Moreover, with a fixed ARL_0, the values of η_0 for the F- and RD-MCUSUM charts are very close. In addition, for the LR based charts, since $\Omega_0 = \Omega_{c,r}$, the second term in the right hand side of (12) is equal to zero, and for the T^2 statistic based charts, it is small compared to the first term. Thus, when we fix ARL_0 to compare ARL_1 between the

Table 3: Drift and variance parameters of the T^2-statistic based charts.

	T^2-F-MCUSUM
d_0	$-k[2\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+4\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+2\tilde{p}]^{1/2}$
$d_{c,r}$	$\tilde{\boldsymbol{\mu}}_{c,r}^{'}(\boldsymbol{\Sigma}_{11}^{-1}+\boldsymbol{\Sigma}_{A^*}^{-1})\tilde{\boldsymbol{\mu}}_{c,r}-k[2\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+4\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+2\tilde{p}]^{1/2}$
Ω_0^2	$2\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+4\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+2\tilde{p}$
$\Omega_{c,r}^2$	$\Omega_0^2+4\tilde{\boldsymbol{\mu}}_{c,r}^{'}(\boldsymbol{\Sigma}_{11}^{-1}+\boldsymbol{\Sigma}_{A^*}^{-1})\boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{11}^{-1}+\boldsymbol{\Sigma}_{A^*}^{-1})\tilde{\boldsymbol{\mu}}_{c,r}$
	T^2-RD-MCUSUM
d_0	$-k(2\tilde{p})^{1/2}$
$d_{c,r}$	$\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}-k(2\tilde{p})^{1/2}$
Ω_0^2	$2\tilde{p}$
$\Omega_{c,r}^2$	$2\tilde{p}+4\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}$

F- and RD-MCUSUM charts, we may compare the values of $\left|\Omega_0^2/(d_0 d_{c,r})\right|$, which we call the *$ARL_1$ measure*.

In the following, denote the ARL_1 measure for LR-F-MCUSUM and LR-RD-MCUSUM charts as m_{LR} and $\tilde{m}_{\text{LR}}$, respectively; and for T^2-F-MCUSUM and T^2-RD-MCUSUM charts as m_{T^2} and $\tilde{m}_{T^2}$, respectively. For LR based charts, we can obtain,

$$m_{\text{LR}}=\frac{4}{\boldsymbol{\mu}_{c,r}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{c,r}}=\frac{4}{\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}+\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{A^*}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}}\leq\frac{4}{\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}}=\tilde{m}_{\text{LR}}. \tag{13}$$

Equation (13) shows that m_{LR} is always smaller than or equal to $\tilde{m}_{\text{LR}}$, which, in turn, implies that ARL_1 of the method using full observation vectors is always smaller than that of the chart with reduced dimension vectors. As a smaller ARL_1 measure implies a smaller ARL_1, we expect that the LR-F-MCUSUM chart generally detects a shift faster than the LR-RD-MCUSUM chart.

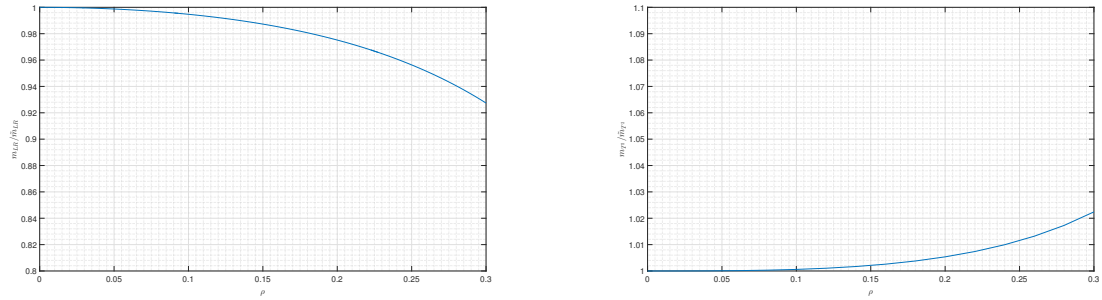
For T^2 based charts, we have

$$m_{T^2}=\left[k^2-k\frac{\tilde{\boldsymbol{\mu}}_{c,r}^{'}(\boldsymbol{\Sigma}_{11}^{-1}+\boldsymbol{\Sigma}_{A^*}^{-1})\tilde{\boldsymbol{\mu}}_{c,r}}{[2\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+4\text{tr}\{\boldsymbol{\Sigma}_{A^*}^{-1}\boldsymbol{\Sigma}_{11}\}+2\tilde{p}]^{1/2}}\right]^{-1}$$

and

$$\tilde{m}_{T^2}=\left[k^2-k\frac{\tilde{\boldsymbol{\mu}}_{c,r}^{'}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\mu}}_{c,r}}{\sqrt{2\tilde{p}}}\right]^{-1}.$$

Note that the theoretical performance measure and the above analysis are applicable to spatial covariance matrix, $\boldsymbol{\Sigma}$, with a general structure. Several commonly used spatial



(a) (b)

Figure 2: Example with $p = 5$, $\tilde{p} = 2$, and $\mu_{c,r} = [1, 1, 0, 0, 0]'$: (a) $m_{\text{LR}}/\tilde{m}_{\text{LR}}$ as a function of ρ; and (b) $m_{T^2}/\tilde{m}_{T^2}$ as a function of ρ.

covariance structures are as follows. In the following, d denotes the distance between two sensors, $C(d|\rho)$ denotes the correlation function between two sensors, which is a function of d and some parameters. Below, $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function of an event, which takes value 1 only when the event is true.

1. Spherical model: $C(d|\rho) = 1\mathbb{1}_{\{d=0\}} + \rho\mathbb{1}_{\{d=1\}} + \frac{\rho}{2}\mathbb{1}_{\{d=\sqrt{2}\}}$ for $\rho \in [0, 1]$.

2. Polynomial model: $C(d|\rho) = 1\mathbb{1}_{\{d=0\}} + \rho^d\mathbb{1}_{\{d>0\}}$ for $\rho \in [0, 1]$.

3. Matérn model: $C(d|\theta) = 1\mathbb{1}_{\{d=0\}} + \frac{1}{2^{v-1}\Gamma(v)}(\sqrt{2}v^{1/2}d/\theta)^v K_v(\sqrt{2}v^{1/2}d/\theta)\mathbb{1}_{\{d>0\}}$ for $\theta > 0$ where K_v is the modified Bessel function of order v (See [46]).

Based on the correlation function, the entries of the covariance matrix $[\mathbf{\Sigma}]_{i,j}$ is determined as $C(d(q_i, q_j)|\theta)$, where q_i and q_j are the coordinates of sensors i and j, respectively.

2.3.2.1 An illustrative example

Using a simple illustrative example, we calculate how much we lose or gain in terms of ARL_1 when reduced dimension vectors are used. Although the ARL_1 measure is applicable to general spatial correlation structure, we use a tridiagonal spatial covariance matrix as an example, which can be regarded as a special case of the spherical model when sensors are located on a 1-dimensional uniform integer grid. The correlation between two sensors is ρ if they are neighboring to each other and 0 otherwise. Such a covariance matrix is denoted by $\mathbf{\Sigma}_1(\rho) \in R_{p\times p}$ with $[\mathbf{\Sigma}_1(\rho)]_{i,j} = 1$ if $i = j$; $[\mathbf{\Sigma}_1(\rho)]_{i,j} = \rho$, if $|i - j| = 1$ and $[\mathbf{\Sigma}_1(\rho)]_{i,j} = 0$, otherwise. We use $p = 5$, $\tilde{p} = 2$ and $\mu_{c,r} = [1, 1, 0, 0, 0]'$ in this example. The

ratio $m_{\text{LR}}/\tilde{m}_{\text{LR}}$ is calculated as a function of the spatial correlation ρ. As spatial correlation is unlikely to be very large in practice, we consider ρ in the range of $0 \leq \rho \leq 0.3$. A ratio smaller than one implies that the charts with full observation vectors have smaller ARL_1 than the charts with reduced observation vectors, and vice versa.

For LR charts, the ratio of ARL_1 measure for the full and reduced-dimension methods can be simplified to

$$\frac{m_{\text{LR}}}{\tilde{m}_{\text{LR}}} = \left[1 + \frac{\tilde{\boldsymbol{\mu}}_{c,r}' \boldsymbol{\Sigma}_{A^*}^{-1} \tilde{\boldsymbol{\mu}}_{c,r}}{\tilde{\boldsymbol{\mu}}_{c,r}' \boldsymbol{\Sigma}_{11}^{-1} \tilde{\boldsymbol{\mu}}_{c,r}}\right]^{-1} = \frac{2 - 6\rho^2}{2 - 5\rho^2 + 3\rho^4}.$$

We plot this ratio as a function of ρ in Figure 2(a). For LR-based charts, the ratio is always smaller than 1 as expected and it decreases as the spatial correlation ρ increases. This indicates that for LR charts, methods based on full observation vectors are always better. However, the performance loss of using reduced-dimensional vectors is small (less than 7% as shown in the plot). Thus, we expect that when the spatial correlation decays reasonably fast, in the case of known center and radius, the reduced-dimension charts do not lose much detection power compared to its full version.

For T^2 statistic based charts, there is no simple analytic expression for the ratio of ARL_1 measures. However, we may still find the ratio numerically. Figure 2(b) shows the plot of $m_{T^2}/\tilde{m}_{T^2}$ as a function of ρ. Interestingly, the ratio is slightly greater than 1, indicating that the reduced dimension method may perform slightly better than their full dimension counterparts, for the particular covariance structure we consider in this example. However, in general, the ratio is smaller than 1 depending on the covariance structure, which we show in Section 2.4.

We also conduct simulation experiments assuming the shift cluster is known. Due to the brevity of the chapter, the results are presented in Appendix A.2. In summary, simulation results match the ARL_1 measures very well. In addition, the RD-MCUSUM charts use only $3 \sim 4\%$ more ARL_1 than the F-MCUSUM charts for a reasonable range of spatial correlation when only a single shift cluster is considered. For T^2 charts, there is almost no performance loss for the settings we considered.

2.4 Experiments

In this section, we conduct numerical experiments under more realistic settings where the shift center and the shift radius are unknown. Then we compare the ARL_1 performance of the F- and the RD-MCUSUM charts using scan statistics.

2.4.1 Experimental setup

We consider the case in which both the shift center c and radius r are unknown. We scan over a set R of possible radii at every possible shift center. In the experiments, we use $R = \{1, \sqrt{2}\}$. In a bell-shaped signal case, we consider $R = \{0, 1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}\}$.

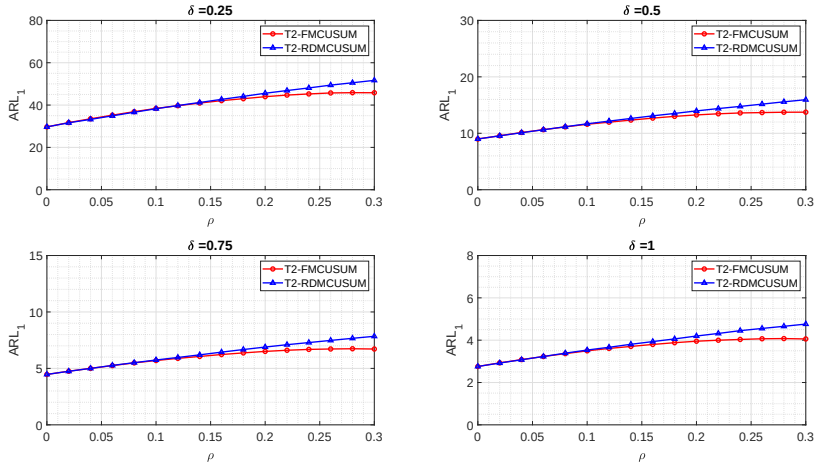
We run the control charts on three commonly used spatial models (spherical model, polynomial model and Matérn model, as introduced in Section 2.3.2). In many aplications such as environmental monitoring, sensors tend to be placed with a reasonable distance and the spatial correlation coefficient between two locations is usually not high. Thus we test $\rho \in \{0,\ 0.02,\ 0.04, \ldots,\ 0.3\}$. In the Matérn model, we test $\theta \in \{0,\ 0.054,\ 0.108, \ldots,\ 0.81\}$ and use order $v = \frac{1}{2}$. Note that $\theta = 0.8$ for the Matérn model corresponds to the spatial correlation among neighboring regions $\rho \approx 0.3$.

The monitored area in the simulation experiments has dimensionality $p = 7 \times 7$. For the out-of-control state, homogeneous shifts (shifts of all affected locations in the cluster have same magnitude) with magnitudes $\delta = 0.25,\ 0.5,\ 0.75$ and 1 are tested. The targeted ARL_0 is fixed to 1000 in all the cases. All simulated ARL values are obtained based on 10,000 simulation replications.

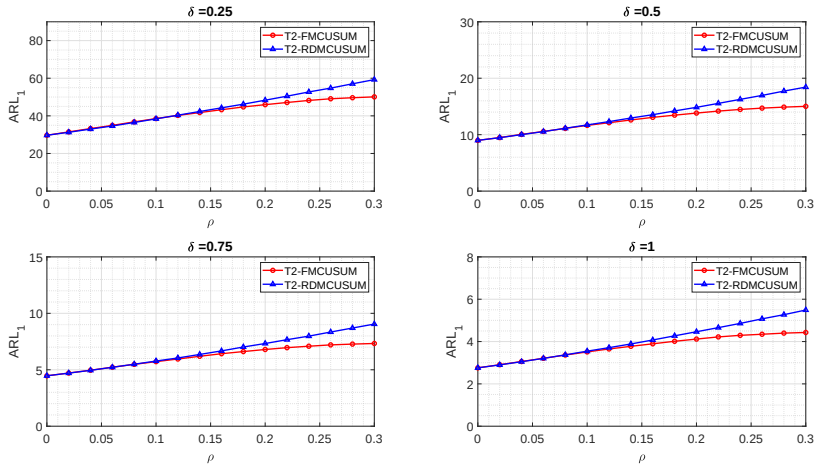
2.4.2 Results

Denote the actual radius as r_{out}. At each time step, the control chart scans over $2p$ possible shift clusters. Figures 3 compares ARL_1 of LR-F-MCUSUM and LR-RD-MCUSUM charts on three different spatial correlation structures. Figures 4 presents results for T^2-F-MCUSUM and T^2-RD-MCUSUM charts.

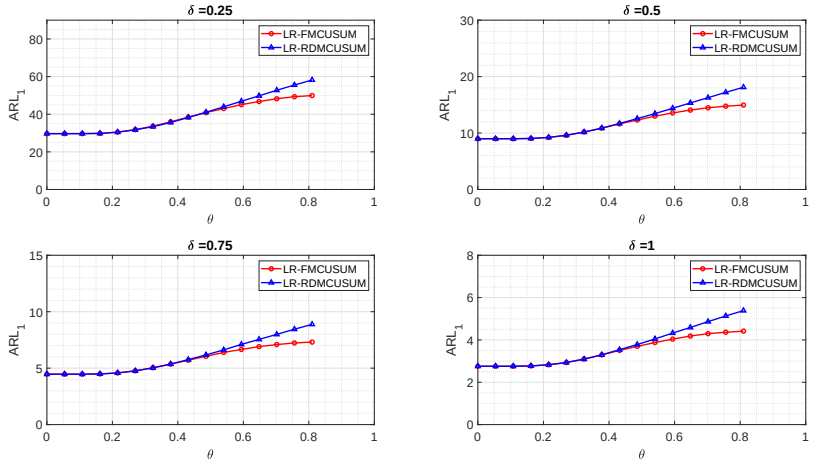
According to the numerical results, we can conclude that in general, the reduced-dimension charts do not severely sacrifice the ARL_1 performance in the range of correlation



(a)

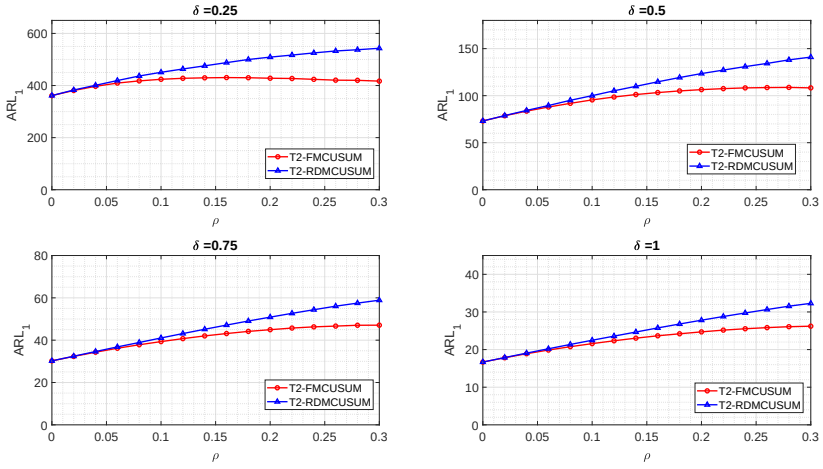


(b)

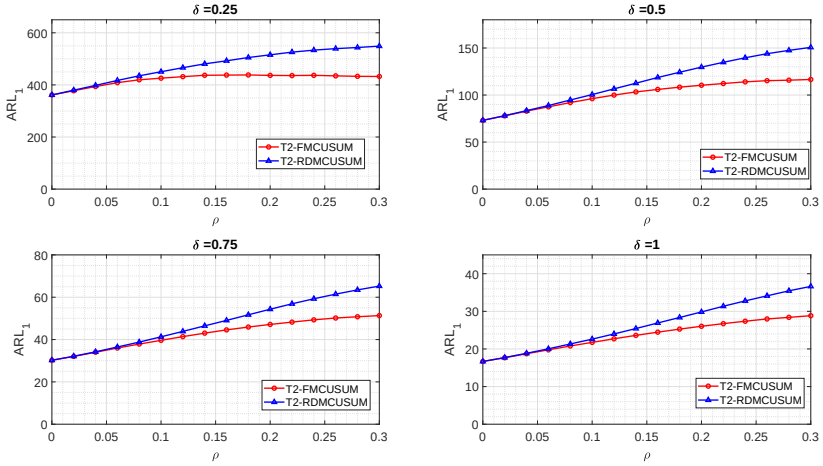


(c)

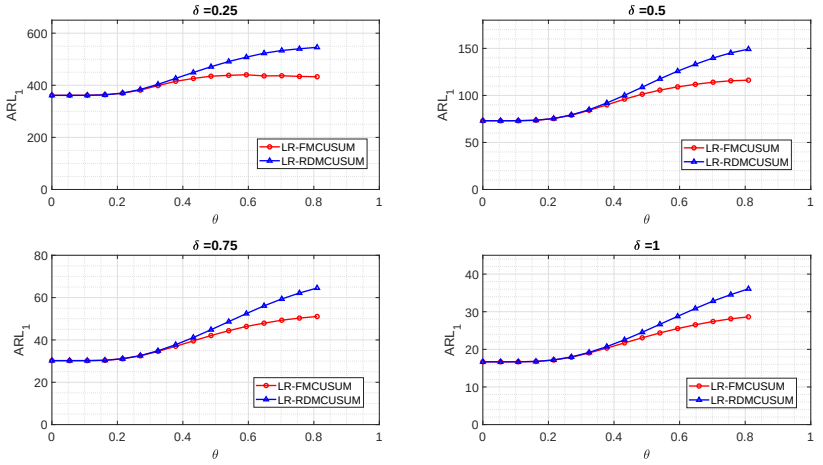
Figure 3: Simulated ARL_1 of LR-F-MCUSUM and LR-RD-MCUSUM charts with $r_{\text{out}} = \sqrt{2}$: (a) spherical model, (b) polynomial model and (c) Matérn model.



(a)



(b)



(c)

Figure 4: Simulated ARL_1 of T^2-F-MCUSUM and T^2-RD-MCUSUM charts with $r_{\mathrm{out}} = \sqrt{2}$: (a) spherical model, (b) polynomial model and (c) Matérn model.

tested here. Thus, the RD-MCUSUM charts can be a powerful and easy-to-implement alternative of the F-MCUSUM charts, especially when the dimension of the monitored area is high and the full covariance matrix is ill-conditioned.

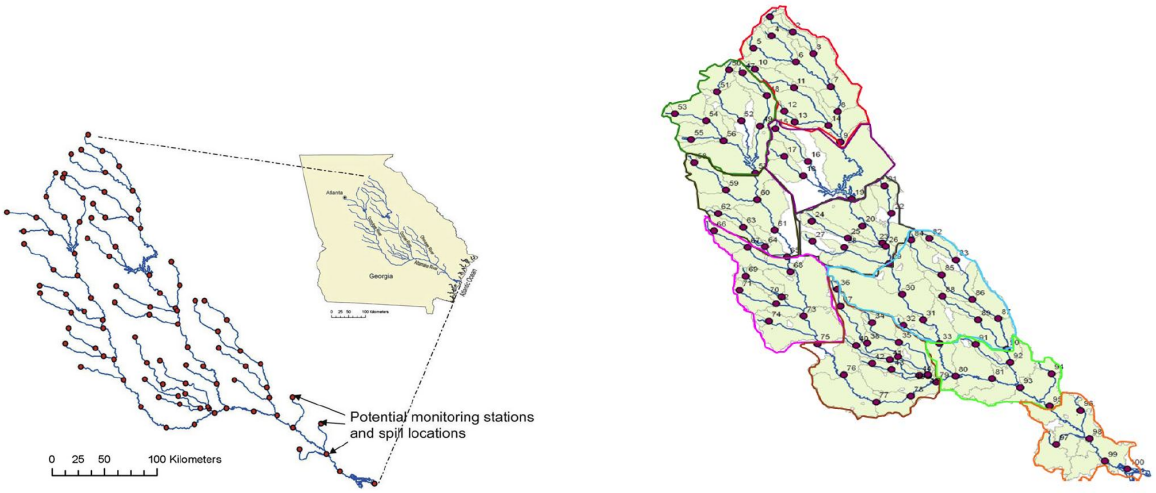
We also consider the case where the signal has a "bell" shaped rather than a boxed shape. In this example, we consider a larger set of possible shift radii, $R = \{0, 1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}\}$. Appendix A.3 includes the performances of the proposed charts for this case and shows how our ARL_1 measure can be used in choosing a scanning radius in the reduced dimension charts.

2.5 Application: Water Quality Monitoring

In this section, we apply the proposed methods to real-time water quality monitoring for a river network. The goal is to detect a contaminant spill that causes water pollution in the river. In this application, the shape of the monitored region is not rectangular and scan clusters are non-circular.

2.5.1 Data

We study the Altamaha River in Georgia, United States, which is the largest watershed in the state. Figure 5 shows the shape of the Altamaha River network with 100 nodes. Each node represents a potential location to place a sensor and also a possible location of a contaminant spill. The contaminant concentration data for such a river network is simulated by the Storm Water Management Model (SWMM) developed by the United States Environmental Protection Agency. SWMM requires geologic, geometric and fundamental hydrodynamics data to construct a river network. In the simulation, rain events and spill events bring randomness to the contaminant transport. Given rainfall information, as well as the location, starting time, intensity and duration of a contaminant spill, SWMM simulates the contaminant transport process through the river over a period of time. We construct the Altamaha river system in the SWMM model based on the United States Geological Survey (USGS) digital elevation data in the National Elevation Dataset [67]. Random rain events are generated based on the patterns obtained in [65]. The Altamaha River watershed is divided into ten sub-catchments as shown in Figure 5(b). The rainfall measurements are



(a) (b)

Figure 5: Shape of the Altamaha River ([65]).

obtained from different USGS stations close to the ten sub-catchments in 2006. Based on the statistical analysis of these measurements, over nine million rain patterns are generated for the entire watershed. Each rain pattern describes time-dependent rainfall events and keeps changing hydrologic conditions in each-catchment during the simulation. For each spill event, one rain pattern is randomly selected and is used in the simulation.

2.5.2 Spatial Models

Due to the nature of hydrodynamics, there exists a spatial correlation among the data streams collected at different locations in the river network. The shape of the network and direction of the stream impose constraints on modeling such spatial correlation. For example, there should not be any correlation for data collected at two monitoring locations that do not share flowing water. A reasonable spatial correlation model is critical to the detection task.

We adopt the so-called "tail-up" spatial model for stream networks, which is proposed based on the moving average constructions in [69]. The tail-up models have the following desired properties: (i) they use stream distance rather than the Euclidean distance, which is defined as the shortest distance along the stream network between two locations; (ii) statistical independence is imposed on the samples located on stream segments that do not

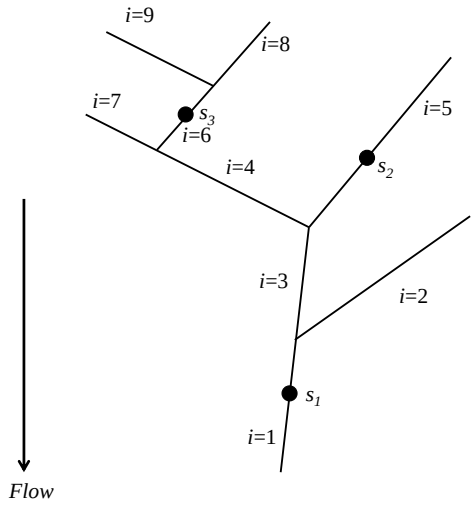


Figure 6: A stream network example with nine stream segments ($i = 1, \ldots, 9$) and three monitoring locations s_1, s_2, s_3.

share a common flow; (iii) proper weighting is incorporated on the entries of the covariance matrix when the line segments in the network is splitting into multiple segments to ensure that the resulting covariance is stationary.

To explain the tail-up model, we first introduce some notations. Suppose a stream network consists of a finite number of stream segments and we index them with $i = 1, 2, \cdots$. Denote the whole set of stream segment indices as I, and the locations on the network as s_j, $j = 1, 2, \cdots$. Let $D_{s_j} \subseteq I$ be the index set of all stream segments that are downstream of location s_j, (which means water from s_j flows into these segments), including the segment containing s_j. Figure 6 illustrates a simple stream network with $I = \{1, 2, \cdots, 9\}$, $D_{s_1} = \{1\}$, $D_{s_2} = \{1, 3, 5\}$ and $D_{s_3} = \{1, 3, 4, 6\}$. Two locations, s_j and s_k are said to be "flow-connected" if $D_{s_j} \cap D_{s_k} = D_{s_j}$ or D_{s_k}. Finally, define

$$B_{s_j,s_k} = \begin{cases} \overline{(D_{s_j} \cap D_{s_k})} \cap (D_{s_j} \cup D_{s_k}), & \text{if } s_j \text{ and } s_k \text{ are flow-connected;} \\ \emptyset, & \text{otherwise.} \end{cases}$$

Here B_{s_j,s_k} represents the set of stream segments between two monitoring locations, including the segment for the upstream location but excluding the one for the downstream location. For example, in Figure 6, $B_{s_1,s_3} = \{3, 4, 6\}$ and $B_{s_2,s_3} = \emptyset$.

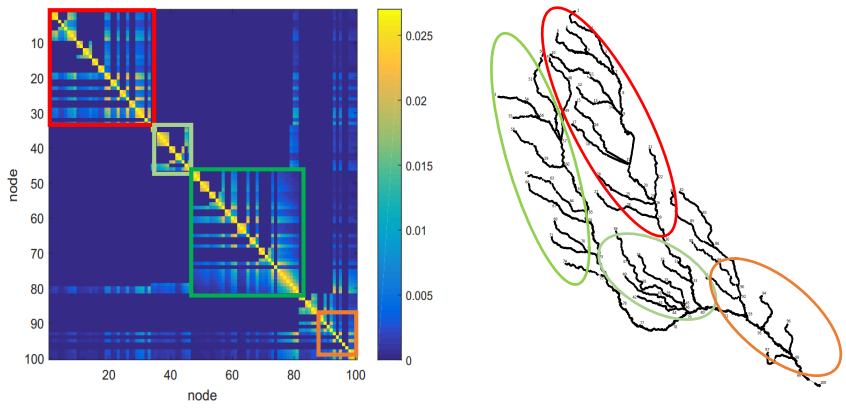


Figure 7: Visualization of the spatial covariance matrix for the Altamaha River. Each block in the covariance matrix corresponds to a branch of the river with a matching color.

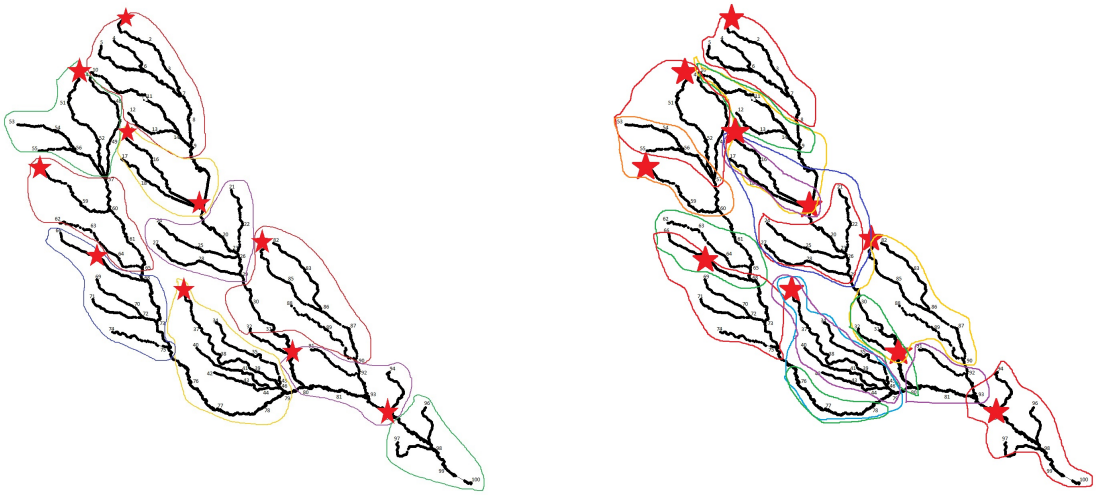
To ensure the stationarity of variances, [69] suggests assigning weights to each stream segment in the network. In a stream network, one segment splits into two segments when it goes up-stream, e.g., in Figure 6, segment 1 splits into segments 2 and 3. One way to weight the segments is based on the flow volume of each segment. For example, we weight segments 2 and 3 by w_2 and w_3, where $w_2 + w_3 = 1$ and w_2/w_3 is equal to the ratio of the flow volume between segments 2 and 3. Using tail-up models, the covariance between two locations, s_j and s_k on the stream network is given by

$$C(s_j, s_k|\boldsymbol{\zeta}) = \begin{cases} 0, & \text{if } s_j \text{ and } s_k \text{ are not flow-connected;} \\ \zeta_1, & \text{if } s_j = s_k; \\ \prod_{i \in B_{s_j,s_k}} \sqrt{w_i}\zeta_1 \rho\big(d(s_j, s_k)/\zeta_2\big), & \text{otherwise;} \end{cases} \tag{14}$$

where $d(s_j, s_k)$ is the stream distance between s_j and s_k, ζ_1 is the variance parameter, $\rho(\cdot|\zeta_2)$ is the correlation function with a parameter ζ_2, and w_i is the weight on the segment i. The correlation function $\rho(\cdot|\zeta_2)$ can be derived from many commonly used spatial models. For illustration, consider the example in Figure 6. If an exponential model is used for spatial correlation, the covariance matrix of s_1, s_2 and s_3 can be constructed as follows:

$$\begin{pmatrix} 1 & \sqrt{w_3 w_5} & \sqrt{w_3 w_4 w_6} \\ \sqrt{w_3 w_5} & 1 & 0 \\ \sqrt{w_3 w_4 w_6} & 0 & 1 \end{pmatrix} \odot \begin{pmatrix} \zeta_1 & \zeta_1 e^{-d(s_1,s_2)/\zeta_2} & \zeta_1 e^{-d(s_1,s_3)/\zeta_2} \\ \zeta_1 e^{-d(s_1,s_2)/\zeta_2} & \zeta_1 & \zeta_1 e^{-d(s_2,s_3)/\zeta_2} \\ \zeta_1 e^{-d(s_1,s_3)/\zeta_2} & \zeta_1 e^{-d(s_2,s_3)/\zeta_2} & \zeta_1 \end{pmatrix},$$

where $\odot$ denotes the Hadamard (element-wise) product operation between two matrices.



(a) (b)

Figure 8: Two sets of scan clusters for spatial scanning: (a) non-overlapped clusters; (b) overlapped clusters. Red stars represent possible spill locations.

We use the tail-up model with an exponential correlation function to model the data collected at different nodes on the Altamaha river network. The covariance matrix for 100 nodes is constructed based on stream distances and flow volume information. We use SWMM to generate in-control data and obtain the maximum likelihood estimator of the parameters in the model, $\hat{\zeta}_1 = 0.027$ and $\hat{\zeta}_2 = 0.68$. The spatial covariance matrix is visualized in Figure 7.

2.5.3 Results

We present the detection performance of the LR-F-MCUSUM and LR-RD-MCUSUM charts for online detection of contaminant spills in the Altamaha river network. Among the 100 nodes on the river network, 10 of them are used as potential contaminant spill locations, which are marked as red stars in Figure 8, and the rest 90 nodes are used for collecting measurements every 15 minutes. In each replication, we run SWMM to simulate the river network during a 10-day period. A single instantaneous spill with a spill location randomly selected from one of the ten possible locations is generated. The spill starting time is uniformly distributed between 15 and 20 hours. Intensity of the contaminant spills follows a uniform distribution, and we consider three different levels: U(10, 100) gram/liter (low), U(100, 250) gram/liter (medium), and U(250, 500) gram/liter (high). Since the Altamaha

Table 4: Detection performance (ARL_1) obtained using the non-overlapped (NOV) and overlapped (OV) sets of scan clusters. Numbers in parentheses are standard errors.

		ARL		False alarm		Fail to detect	
	Intensity	F	RD	F	RD	F	RD
NOV	low	40.68 (4.11)	46.91 (4.60)	7%	7%	8%	9%
	medium	30.11 (2.31)	33.45 (2.90)	7%	7%	2%	9%
	high	23.79 (1.39)	25.65 (1.68)	9%	7%	0%	0%
OV	low	36.45 (3.73)	39.16 (4.22)	8%	9%	4%	8%
	medium	33.09 (2.54)	30.44 (2.26)	4%	3%	1%	3%
	high	27.30 (2.76)	27.68 (2.43)	4%	3%	1%	3%

river network does not have a regular shape and the monitoring stations are not located on the uniform grid, we do not use the circle shape clusters for spatial scanning. Instead, we construct scan clusters based on locations of the sensors and topology of the river network. Two sets of clusters are considered: (i) ten clusters that have no overlap (Figure 8(a)); and (ii) 18 clusters with partial overlap (Figure 8(b)). For the LR based MCUSUM charts, we set the minimum size of change that we aim to detect to be 0.05 gram/liter and use it to construct the post-change mean vectors, $\boldsymbol{\mu}_{c,r}$ and $\tilde{\boldsymbol{\mu}}_{c,r}$. To have a fair comparison, the thresholds for both LR-F-MCUSUM and LR-RD-MCUSUM charts are adjusted so that the in-control average run lengths are 10 days (960 samples). We generate 300 simulated contaminant spills (100 spills in each level of intensity). Detection performances of the two methods using the two sets of scan clusters are summarized in Table 4. From Table 4, we can see that the ARL performances of the two methods are consistent with our analysis in the previous sections: the chart using full observations achieves slightly smaller ARL but it is possible that the RD chart performs better as in the overlapping-medium case. To further compare the performance between the LR-F-MCUSUM and LR-RD-MCUSUM charts, we also calculate the percentage of cases where (i) the chart using full observations performs better; (ii) the two methods perform similarly, i.e., both methods successfully detect a spill event and the absolute difference of detection delays between the two methods is less than 1 hour; (iii) the chart using RD observations performs better; and (iv) both methods fail to detect by either missing the spill or raising a false alarm. These results are presented in Table 5. The table shows that, almost under all settings, the percentage

Table 5: Detection delay comparison between LR-F-MCUSUM and LR-RD-MCUSUM charts.

	non-overlapped			overlapped		
	low	medium	high	low	medium	high
Full better	35%	34%	34%	30%	18%	27%
Similar	34%	43%	45%	40%	58%	54%
RD better	25%	19%	20%	26%	22%	19%
Both fail	6%	4%	1%	4%	2%	0%

of cases where the LR-RD-MCUSUM chart performs no worse than (performs better than or similarly with) LR-F-MCUSUM chart is higher than 60%. Hence, even if the charts using full observations achieve slightly smaller ARL values, the RD based charts show very competitive performances in practice. Given that the RD based charts enjoy distributed computing with a little loss in ARL performances, they should be considered as a good alternative or even a better option for large-scale sensor networks especially when the full covariance matrix is ill-conditioned.

2.6 Conclusion

In this chapter, we study the reduced-dimension control charts for spatial-temporal change-point detection in the presence of various spatial covariance structures. The reduced-dimension charts perform spatial scanning by breaking the entire monitoring area into overlapping clusters, as well as computing the detection statistics locally while incorporating local covariance. In the presence of high-dimensional data streams, the reduced-dimension approach enjoys lower communication complexity and better numerical stability. We quantify the performance loss or gain due to dimension reduction through systematic theoretical and numerical studies. Considering the benefits of dimensionality reduction, the reduced dimension charts can be a powerful and cheaper alternative to the full-observations charts in high-dimensional problems.

CHAPTER III

$\mathsf{S}^3\mathsf{T}$: A SCORE STATISTIC FOR SPATIO-TEMPORAL CHANGE-POINT DETECTION

Detection of the emergence of a signal in a noisy background arises in many multi-sensor spatio-temporal change-point detection applications. When the monitored process is in-control, sensors observe noise. When the monitored process is out of control, a signal emerges in the noise. A variety of applications possess particular spatial and temporal correlation structures. One application is an environmental sensor network, which is used to monitor river systems to detect a contaminant hazard [26]. When the signal emerges, observations from sensors may have a time-varying mean and spatio-temporal correlation structures due to water flow.

Exploiting spatio-temporal structures of the change is crucial for detecting weaker signals. However, most existing methods only capture either spatial correlations [19, 11, 22, 30, 29] or temporal correlations [73]. It is still unclear how to jointly capture the spatial and temporal correlations in detection statistics. Moreover, computational complexity is often a concern when we try to jointly model spatial and temporal correlation, especially when there are a large number of sensors that lead to high-dimensional observations. In particular, one issue with the likelihood ratio statistic is that one has to invert the sample covariance matrix, which can be computationally expensive and numerically unstable. An alternative to the likelihood ratio statistic is the score statistic, which can sometimes lead to a simpler test statistic. When the hypothesis involves a univariate parameter, the score test is the locally most powerful test [44].

In this chapter, we propose a new efficient score statistic for spatial-temporal change-point detection, which we call the $\mathsf{S}^3\mathsf{T}$ statistic. The $\mathsf{S}^3\mathsf{T}$ statistic can capture both spatial and temporal correlations of the signal. Hence, it can react quickly to a change in the mean and/or the spatio-temporal covariance. The score statistic is computationally efficient. By

avoiding the inversion of the sample covariance matrix, the $\mathsf{S}^3\mathsf{T}$ statistic has computation complexity $O(p^3)$, where p is the dimensionality of the observations, whereas the likelihood ratio statistic has $O(p^3N^3)$ complexity, which grows with the time horizon N. Our main theoretical contributions are analytic approximations for the false alarm rate in the offline case and the in-control average run length in the online case. The theoretical approximations can be used for calibrating thresholds to control the false alarm rate of our procedure. For scalar observations, our statistic $\mathsf{S}^3\mathsf{T}$ reduces to the score detector considered in [73]. Our work provides a novel extension of [73] for multi-dimensional observations when there are both spatial and temporal correlations.

The rest of the chapter is organized as follows. Section 3.1 formulates the problem. Section 3.2 presents detection statistics for both offline and online change-point detection. Section 3.3 presents theoretical approximations for the false alarm rate in the offline case and in-control average-run-length in the online case. Section 3.4 contains numerical examples for simulated data and real data, as well as a case study of water quality monitoring. Finally, Section 3.5 concludes the chapter.

3.1 Problem Formulation

Consider a sequence of samples $\boldsymbol{y}_\ell \in \mathbb{R}^p$, $\ell = 1, 2, \cdots, N$, where p is the dimension, N is the sample size, which is fixed in the offline setting and grows in the online setting. We assume that under the null hypothesis, $\{\boldsymbol{y}_\ell\}$ forms a series of i.i.d. normal random vectors with spatial correlation caused by, for instance, sensor measurement errors or background noises from the environment. At an unknown time k, $1 \leq k \leq N-1$, which corresponds to the unknown change-point, a signal emerges on top of the noise. The change may alter not only the mean of $\{\boldsymbol{y}_\ell\}$ but also the spatio-temporal correlation structure, which we will explain in more details.

First consider an offline setting, where the goal is to detect a change in retrospect from

the samples. Formally, this can be formulated as the following hypothesis test:

$$\begin{aligned} \mathsf{H}_0: &\quad \boldsymbol{y}_\ell = \boldsymbol{w}_\ell, && \ell = 1, 2, \cdots, N, \\ \mathsf{H}_1: &\begin{cases} \boldsymbol{y}_\ell = \boldsymbol{w}_\ell, & \ell = 1, 2, \cdots, k, \\ \boldsymbol{y}_\ell = \boldsymbol{x}_\ell + \boldsymbol{w}_\ell, & \ell = k+1, \cdots, N, \end{cases} \end{aligned} \tag{15}$$

where $\boldsymbol{w}_\ell \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is the spatial covariance matrix of the noise. We assume that before the change, the samples have no temporal correlation. This is reasonable, because we often have enough reference data before the change to estimate and then remove the temporal correlation.

Below we describe a model for the signal $\{\boldsymbol{x}_\ell\}$ after the change has occurred. The signal can be spatially and temporally correlated. We capture the temporal correlation using multivariate time-series models. Two examples are the first-order vector autoregressive VAR(1) model [5],

$$\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta \boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell, \quad \ell = 1, 2, \ldots, \tag{16}$$

where $\theta \in \mathbb{R}$, $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}_\ell]$ and $\boldsymbol{\epsilon}_\ell$ is the process noise; and the VARMA(1, 1) model given by

$$\boldsymbol{x}_{\ell+1} + \phi \boldsymbol{x}_\ell = (1 + \phi - \eta)\boldsymbol{\mu} + \eta \boldsymbol{\epsilon}_\ell + \boldsymbol{\epsilon}_{\ell+1}, \quad \ell = 1, 2, \ldots,$$

where the parameters are $\eta \in \mathbb{R}$ and $\phi \in \mathbb{R}$. Models with higher orders can also be used if necessary.

We capture the spatial correlation of the signal using standard spatial correlation models [13]. Denote $\mathrm{Var}[\boldsymbol{x}_\ell] = \gamma \boldsymbol{\Lambda} \in \mathbb{R}^{p \times p}$, where $\boldsymbol{\Lambda}$ is the spatial correlation matrix of the signal $\boldsymbol{x}_\ell$, and $\gamma \geq 0$ is the magnitude (assuming the model is stationary and $\mathrm{Var}[\boldsymbol{x}_\ell]$ does not change over ℓ). Note that the variance of the signal $\mathrm{Var}[\boldsymbol{x}_\ell]$ depends on the variance of the process noise, $\mathrm{Var}[\boldsymbol{\epsilon}_\ell]$. Here we assume the *structure* of $\boldsymbol{\Lambda}$ is known but the parameter γ may be unknown. This is a common practice, because once a spatial correlation model is assumed, $\boldsymbol{\Lambda}$ is usually specified by the location of the samples and some unknown parameters. In particular, each entry of the spatial covariance $\boldsymbol{\Lambda}$ is determined by a correlation function, $C(d|\rho)$, of the distance d between two samples (sensors) and is parameterized by ρ. Let $\mathbb{1}\{A\}$ denotes an indicator function, which takes value 1 when the event A is true, and 0 otherwise. Several commonly used correlation functions are:

(i) Spherical model [30]:

$$C(d|\rho) = 1\mathbb{1}\{d=0\} + \rho\mathbb{1}\{d=1\} + \frac{\rho}{2}\mathbb{1}\{d=\sqrt{2}\}, \quad \rho \in [0,1]; \tag{17}$$

(ii) Exponential model [13]:

$$C(d|\rho) = 1\mathbb{1}\{d=0\} + e^{-d/\rho}\mathbb{1}\{d>0\}, \quad \rho > 0;$$

(iii) Matérn model [13]:

$$C(d|\rho) = 1\mathbb{1}\{d=0\} + \frac{1}{2^{v-1}\Gamma(v)}(\sqrt{2}v^{1/2}d/\rho)^v K_v(\sqrt{2}v^{1/2}d/\rho)\mathbb{1}\{d>0\}, \quad \rho > 0;$$

where $\rho > 0$, v is the order of the Matérn model that determines the degree of smoothness of the correlation function, $\Gamma(\cdot)$ is the gamma function, $K_v(\cdot)$ is the modified Bessel function of the second kind [46]. Note that when $v = p + 0.5$, $p \in \mathbb{R}^+$, the Matérn model is a product of an exponential and a polynomial of order p. When $v = 0.5$, the Matérn model is equivalent to the exponential model. When $v \to \infty$, it converges to the squared exponential covariance function.

Now we derive our detection statistic. For an assumed change location k, let

$$\tau = N - k$$

denote the number of post-change samples. Define a vector by concatenating all samples after the assumed change-point location k,

$$\boldsymbol{y}_{(k+1:N)} = [\boldsymbol{y}_{k+1}^\intercal, \cdots, \boldsymbol{y}_N^\intercal]^\intercal \in \mathbb{R}^{p\tau}, \tag{18}$$

where $a^\intercal$ denotes the transpose of a vector a. This step is demonstrated in Figure 9. Define $\boldsymbol{x}_{(k+1:N)}$ and $\boldsymbol{w}_{(k+1:N)}$ similarly. Then after the change, we have

$$\boldsymbol{y}_{(k+1:N)} = \boldsymbol{x}_{(k+1:N)} + \boldsymbol{w}_{(k+1:N)}.$$

The covariance matrix of the concatenating observation vector consists of two terms that are due to the signal and the noise, respectively:

$$\mathrm{Var}[\boldsymbol{y}_{(k+1:N)}] = \gamma \boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau,$$

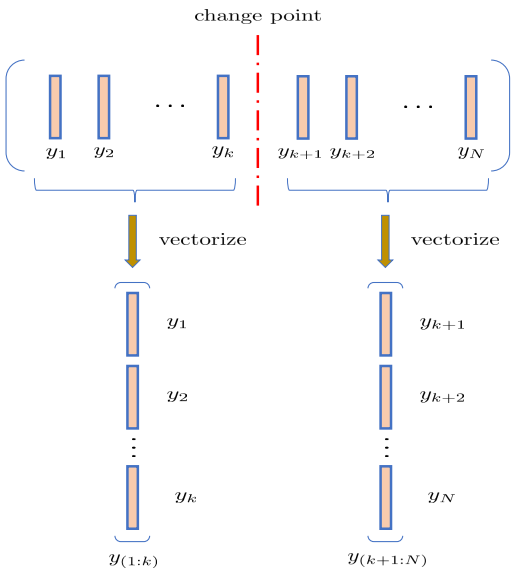


Figure 9: Diagram showing the concatenation of samples.

where $\gamma \boldsymbol{V}_\tau(\theta) = \mathrm{Var}[\boldsymbol{x}_{(k+1:N)}]$, $\boldsymbol{\Sigma}_\tau = \mathrm{Var}[\boldsymbol{w}_{(k+1:N)}]$, and θ is a parameter related to the temporal correlation which we will specify next. The second term in the covariance matrix is given by

$$\boldsymbol{\Sigma}_\tau = \boldsymbol{I}_\tau \otimes \boldsymbol{\Sigma} \in \mathbb{R}^{p\tau \times p\tau}, \tag{19}$$

where $\boldsymbol{I}_\tau$ is a τ-by-τ identity matrix and $\otimes$ denotes the Kronecker product.

By concatenating the observation vectors as in (18), we can jointly model spatial and temporal correlation of the signal by one matrix $\boldsymbol{V}_\tau(\theta)$. For instance, for VAR(1) model,

$$\boldsymbol{V}_\tau(\theta) = \boldsymbol{R}_\tau(\theta) \otimes \boldsymbol{\Lambda}, \tag{20}$$

where $\boldsymbol{R}_\tau(\theta) \in \mathbb{R}^{\tau \times \tau}$ and $[\boldsymbol{R}_\tau(\theta)]_{i,j} = \theta^{|i-j|}, \forall i, j \in \{1, \cdots, \tau\}$ is due to the temporal correlation in (16). Similarly, if the signal follows the VARMA(1,1) model, the matrix $\boldsymbol{V}$ can be parameterized by $\theta \triangleq (\phi, \eta)$ with the following form:

$$\boldsymbol{V}_\tau(\theta) = \boldsymbol{R}_\tau(\phi, \eta) \otimes \boldsymbol{\Lambda}, \tag{21}$$

where $\boldsymbol{R}_\tau(\phi, \eta) \in \mathbb{R}^{\tau \times \tau}$; $[\boldsymbol{R}_\tau(\phi, \eta)]_{i,j} = 1 + \eta^2 - 2\phi\eta$, if $i = j$ and $[\boldsymbol{R}_\tau(\phi, \eta)]_{i,j} = \phi^{|i-j|-1}(\phi - \eta)(1 - \phi\eta)$, otherwise. For other models, similar forms of $\boldsymbol{V}_\tau$ can be derived: the temporal

Table 6: Notations.

p	dimension of samples
N	sample size in offline change-point detection
k	change-point location
τ	number of post change samples, $\tau = N - k$
$\boldsymbol{\Sigma}$	spatial covariance matrix of the noise, $\boldsymbol{\Sigma} = \mathrm{Var}[\boldsymbol{w}_\ell]$
$\boldsymbol{\Lambda}$	structure of spatial covariance matrix of the signal $\mathrm{Var}[\boldsymbol{x}_\ell] = \gamma\boldsymbol{\Lambda}$
γ	magnitude of spatial covariance matrix of the signal $\mathrm{Var}[\boldsymbol{x}_\ell] = \gamma\boldsymbol{\Lambda}$
$\boldsymbol{\Sigma}_\tau$	covariance of noise in concatenated observations $\boldsymbol{\Sigma}_\tau = \mathrm{Var}[\boldsymbol{w}_{(k+1:N)}] = \boldsymbol{I}_\tau \otimes \boldsymbol{\Sigma}$
$\gamma\boldsymbol{V}_\tau(\theta)$	covariance of signal in concatenated observations $\mathrm{Var}[\boldsymbol{x}_{(k+1:N)}] = \gamma\boldsymbol{V}_\tau(\theta) = \gamma\boldsymbol{R}_\tau(\theta) \otimes \boldsymbol{\Lambda}$
$\boldsymbol{R}_\tau(\theta)$	matrix that captures temporal dependence of the signal

dependence of the signal is captured by $\boldsymbol{R}_\tau$, the spatial dependence by $\boldsymbol{\Lambda}$, and the spatial-temporal covariance is a Kronecker product of the two [16].

Using the representation above, the detection problem can be reformulated as the following hypothesis test:

$$\begin{aligned} \mathsf{H}_0: \quad & \boldsymbol{y}_{(1:k)} \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}_k\big), \quad \boldsymbol{y}_{(k+1:N)} \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}_\tau\big), \\ \mathsf{H}_1: \quad & \boldsymbol{y}_{(1:k)} \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}_k\big), \quad \boldsymbol{y}_{(k+1:N)} \sim \mathcal{N}\big(\boldsymbol{\mu}_{(k+1:N)},\ \gamma\boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau\big), \end{aligned} \tag{22}$$

for $k = 1, \cdots, N-1$, where $\boldsymbol{0}$ is a vector of zeros, $\boldsymbol{\mu}_{(k+1:N)} = \mathrm{E}[\boldsymbol{y}_{(k+1:N)}] \in \mathbb{R}^{p\tau}$ and $\gamma \in \mathbb{R} > 0$. Note that we assume $\boldsymbol{\mu}_{(k+1:N)}$ is unknown. Equivalently, under the null hypothesis, $\gamma = 0$ and $\boldsymbol{\mu}_{(k+1:N)} = \boldsymbol{0}$, and under the alternative hypothesis, $\gamma > 0$ or $\boldsymbol{\mu}_{(k+1:N)} \neq \boldsymbol{0}$. Using this form of hypothesis, we can derive our score statistic.

Table 6 provides a list of notations used throughout the chapter.

3.2 Statistic for Offline and Online Detection

We now derive the $\mathsf{S}^3\mathsf{T}$ statistic for offline change-point detection. The log-likelihood function for the hypothesis test in (22) is given by

$$\begin{aligned} \ell(\gamma, \boldsymbol{\mu}, \tau, \theta) = & -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left|\gamma\boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau\right| \\ & - \frac{1}{2}(\boldsymbol{y}_{(k+1:N)} - \boldsymbol{\mu}_{(k+1:N)})^\intercal(\gamma\boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau)^{-1}(\boldsymbol{y}_{(k+1:N)} - \boldsymbol{\mu}_{(k+1:N)}). \end{aligned} \tag{23}$$

To cope with unknown parameters, we may use the generalized likelihood ratio (GLR) statistic based on (23). However, (23) involves the inversion of a $p\tau$-by-$p\tau$ dimensional matrix $\gamma\boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau$, which incurs a complexity of $O(p^3\tau^3)$ for a given τ. Recall that

$\tau = N - k$, so $\tau = 1, 2, \ldots, N$. Hence, the complexity of computing the GLR statistic is $O(p^3N^3)$, which grows polynomially with N (the time horizon), and the computation of the likelihood statistic becomes prohibitive.

3.2.1 Quadratic score statistic

Define the following notations. Let $\boldsymbol{A}_\tau(\theta) = \boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)$, $\boldsymbol{B}_\tau(\theta) = \boldsymbol{\Sigma}_\tau^{-1/2}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1/2}$, $c(\tau,\theta) = \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\big)$, and $d(\tau,\theta) = 2\mathrm{tr}\big[\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big]$, where $\mathrm{tr}\big(\,\cdot\,\big)$ denotes the trace of a matrix. We now derive the score-statistic for detection. Take the derivative of $\ell(\gamma, \boldsymbol{\mu}, \tau, \theta)$ in (23) with respect to γ and $\boldsymbol{\mu}$ and evaluate at $\gamma = 0$ and $\boldsymbol{\mu} = \mathbf{0}$. We can obtain

$$\varsigma(\tau,\theta) = \begin{bmatrix} \frac{\partial \ell}{\partial \gamma}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0} \\ \frac{\partial \ell}{\partial \boldsymbol{\mu}}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}c(\tau,\theta) + \frac{1}{2}\boldsymbol{y}_{(k+1:N)}^\intercal\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}_{(k+1:N)} \\ \boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}_{(k+1:N)} \end{bmatrix}. \tag{24}$$

The derivation of (24) is given in Appendix B.1. It can be verified that $\mathrm{E}[\varsigma(k,\theta)] = \mathbf{0}$ under the null hypothesis, where $\mathbf{0}$ represents the zero vector. It can also be shown that the covariance of the score vector $\varsigma(\tau,\theta)$ is given by

$$\mathrm{Cov}[\varsigma(\tau,\theta)] = \begin{bmatrix} \frac{1}{4}d(\tau,\theta) & 0 \\ \mathbf{0} & \boldsymbol{\Sigma}_\tau^{-1} \end{bmatrix}.$$

As suggested by [43], when the likelihood function involves multiple parameters, the score statistic is a quadratic function of the efficient score vector. In our case, this becomes

$$\begin{aligned} S(\tau,\theta) &= \varsigma(\tau,\theta)^\intercal \mathrm{Cov}[\varsigma(\tau,\theta)]^{-1}\varsigma(\tau,\theta) \\ &= \frac{\Big[\boldsymbol{y}_{(k+1:N)}^\intercal\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}_{(k+1:N)} - c(\tau,\theta)\Big]^2}{d(\tau,\theta)} + \boldsymbol{y}_{(k+1:N)}^\intercal\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}_{(k+1:N)}. \end{aligned} \tag{25}$$

The most expensive part in evaluating (25) is to compute $\boldsymbol{\Sigma}_\tau^{-1}$. According to (19), we have $\boldsymbol{\Sigma}_\tau^{-1} = \boldsymbol{I}_\tau \otimes \boldsymbol{\Sigma}^{-1}$, which means that we only need to compute $\boldsymbol{\Sigma}^{-1}$ that has a complexity $O(p^3)$. Hence, the computation complexity of evaluating $S(\tau,\theta)$ is much lower than that of the GLR statistic. Moreover, since $\boldsymbol{\Sigma}$ is assumed known and fixed, its inversion can be pre-computed; however, in (23), the likelihood function involves $(\gamma\boldsymbol{V}_\tau(\theta) + \boldsymbol{\Sigma}_\tau)^{-1}$, which has to be computed for each τ value.

Since the expected value of $S(\tau,\theta)$ increases as τ increases, it needs to be normalized to have mean 0 and variance 1 under the null hypothesis. This leads to the following *quadratic*

score statistic,

$$\widetilde{S}(\tau,\theta)=\frac{S(\tau,\theta)-\mathrm{E}\big[S(\tau,\theta)\big]}{\sqrt{\mathrm{Var}\big[S(\tau,\theta)\big]}}, \tag{26}$$

where $\mathrm{E}\big[S(\tau,\theta)\big]=p\tau+1$, and the variance is given by (derivation can be found in Appendix B.2)

$$\begin{aligned}\mathrm{Var}\big[S(\tau,\theta)\big]=&2p\tau+10-24\frac{c(\tau,\theta)}{d(\tau,\theta)^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\Big)\\&+\frac{48}{d(\tau,\theta)^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\Big).\end{aligned}$$

Then we may construct the *quadratic detector* using $\widetilde{S}(\tau,\theta)$, which detects a signal when the maximum standardized score statistic over all possible parameter values of $\theta\in\Theta$ and τ exceeds a pre-specified threshold $b>0$,

$$\max_{\theta\in\Theta,\ 1\le\tau\le N}\widetilde{S}(\tau,\theta)\ge b.$$

3.2.2 $\mathsf{S}^3\mathsf{T}$ statistic for offline change-point detection

Although the quadratic score statistic achieves the maximum discrimination between the null and the alternative distribution [43], theoretical analysis of the detection statistic is intractable; thus, it is difficult to calibrate the threshold b to control the false alarm rate. In this section, we propose a simpler statistic, namely the $\mathsf{S}^3\mathsf{T}$ statistic, which is the score statistic with respect to γ *only*:

$$W(\tau,\theta)=\frac{\frac{\partial\ell}{\partial\gamma}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0}}{\sqrt{\mathrm{Var}\big[\frac{\partial\ell}{\partial\gamma}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0}\big]}}=\frac{\boldsymbol{y}_{(N-\tau+1:N)}^{\mathsf{T}}\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}_{(N-\tau+1:N)}-c(\tau,\theta)}{\sqrt{d(\tau,\theta)}}. \tag{27}$$

Note that both the spatial and temporal correlations of the signal are still captured in the statistic by $\boldsymbol{V}_\tau(\theta)$ in (20). Under the null hypothesis, the detection statistic $W(\tau,\theta)$ has mean 0 and unit variance. The detection procedure claims a change when the maximum of the score statistic exceeds a pre-specified threshold $b>0$,

$$\max_{\theta\in\Theta,\ 1\le\tau\le N}W(\tau,\theta)\ge b. \tag{28}$$

3.2.3 $\mathsf{S}^3\mathsf{T}$ statistic for online change-point detection

We now present an online change-point detection procedure based on the $\mathsf{S}^3\mathsf{T}$ statistic. In the online setting, the sample size N is not fixed and samples are sequentially collected. A

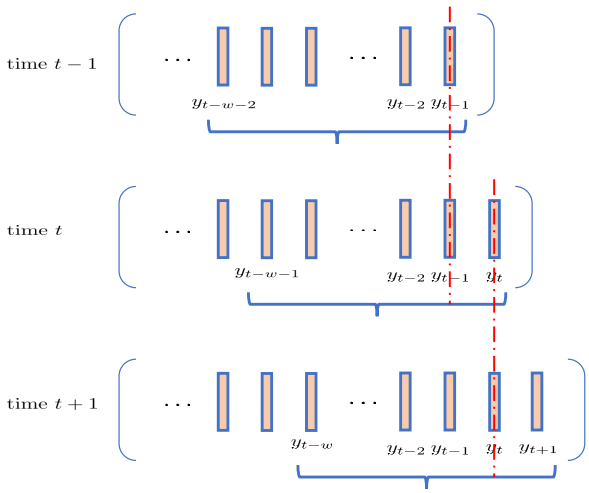


Figure 10: Sliding window of length w for online detection.

signal may occur at an unknown time t. Our goal is to detect the emergence of the signal as soon as possible.

Consider a sequential version of the hypothesis test in (15), where the number of samples N increase. We adopt a sliding window approach for online detection, and construct the detection statistic using the most recent ω samples at each time, where ω is a pre-specified window length (demonstrated in Figure 10). Given a current time t, the detection statistic constructed using the most recent ω samples is given by

$$W_t(\omega, \theta) = \frac{\boldsymbol{y}_{(t-\omega+1:t)}^{\mathsf{T}} \boldsymbol{\Sigma}_\omega^{-1} \boldsymbol{V}_\omega(\theta) \boldsymbol{\Sigma}_\omega^{-1} \boldsymbol{y}_{(t-\omega+1:t)} - c(\omega, \theta)}{\sqrt{d(\omega, \theta)}}. \tag{29}$$

The detection procedure for online change-point detection is a stopping time, which raises an alarm when the detection statistic exceeds a threshold $b > 0$ for the first time:

$$\mathcal{T} = \inf \left\{ t : \max_{\theta \in \Theta} W_t(\omega, \theta) \geq b \right\}. \tag{30}$$

3.3 Theoretical Approximations

3.3.1 Significance level for offline $\mathsf{S}^3\mathsf{T}$ statistic

We present a theoretical approximation for the significance level of the detection procedure defined in (28). The approximation is quite accurate and can be used to avoid the time-consuming simulation when choosing an appropriate b. Denote the standard normal density

function by $\phi(x)$ and its distribution function by $\Phi(x)$, and define a special function [53]:

$$\nu(x) = \frac{\frac{2}{x}\Big[\Phi\big(\frac{x}{2}\big) - \frac{1}{2}\Big]}{\frac{x}{2}\Phi\big(\frac{x}{2}\big) + \phi\big(\frac{x}{2}\big)}. \tag{31}$$

Define the following quantities, which are useful to state our theoretical approximation results:

$$\mu(\tau,\theta) = \tau\left[\frac{\mathrm{tr}\big(\boldsymbol{A}_{\tau+1}(\theta)\boldsymbol{A}_{\tau+1}(\theta)\big)}{\mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)} - 1\right], \tag{32}$$

$$H(\tau,\theta) = -\frac{\partial^2 E[W(\tau,\theta)W(\tau,s)]}{\partial^2 s}\bigg|_{s=\theta}, \tag{33}$$

$$\psi(\xi) = -\xi\frac{c(\tau,\theta)}{\sqrt{d(\tau,\theta)}} - \frac{1}{2}\log\bigg|\boldsymbol{I}_{p\tau} - \frac{2\xi\boldsymbol{B}_{\tau}(\theta)}{\sqrt{d(\tau,\theta)}}\bigg|. \tag{34}$$

Note that $\psi(\xi)$ is the cumulant generating function (a.k.a., the log-moment generating function) of the detection statistic $W(\tau,\theta)$. The following theorem is our main theoretical result, which provides an analytical approximation for the significance level of the detection procedure defined in (28).

Theorem 3.3.1 (Approximation for significance level)**.** *When the threshold $b \to \infty$ and $\theta \in \Theta \subset \mathbb{R}^d$, under the null hypothesis, the probability of false alarm for the procedure defined in* (28) *is given by*

$$\begin{aligned}&\mathbb{P}_{H_0}\Big(\max_{\substack{\theta\in\Theta\\1\le\tau\le N}} W(\tau,\theta) \ge b\Big)\\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\sum_{\tau=1}^{N}\int_{\theta\in\Theta}\frac{[b\xi_0(\tau,\theta)]^{\frac{d}{2}}}{\xi_0(\tau,\theta)}g(\tau,\theta)|H(\tau,\theta)|^{\frac{1}{2}}\frac{b^2\mu(\tau,\theta)}{2\tau}\nu\Big(\sqrt{\frac{b^2\mu(\tau,\theta)}{\tau}}\Big)d\theta + o(1),\end{aligned} \tag{35}$$

where

$$g(\tau,\theta) = \frac{\exp\big(-\xi_0(\tau,\theta)b + \psi(\xi_0(\tau,\theta))\big)}{\sigma_{\xi_0}\sqrt{2\pi}}, \tag{36}$$

$$\sigma^2_{\xi_0} = d(\tau,\theta)^{-1}\mathrm{tr}\Big(\Big[\boldsymbol{I}_{p\tau} - \frac{2\xi_0\boldsymbol{B}_{\tau}(\theta)}{\sqrt{d(\tau,\theta)}}\Big]^{-1}\boldsymbol{B}_{\tau}(\theta)\Big[\boldsymbol{I}_{p\tau} - \frac{2\xi_0\boldsymbol{B}_{\tau}(\theta)}{\sqrt{d(\tau,\theta)}}\Big]^{-1}\boldsymbol{B}_{\tau}(\theta)\Big), \tag{37}$$

and $\xi_0(\tau,\theta)$ is the solution to

$$\frac{1}{\sqrt{d(\tau,\theta)}}\mathrm{tr}\Big(\Big[\boldsymbol{I}_{p\tau} - \frac{2\xi_0\boldsymbol{B}_{\tau}(\theta)}{\sqrt{d(\tau,\theta)}}\Big]^{-1}\boldsymbol{B}_{\tau}(\theta) - \boldsymbol{A}_{\tau}(\theta)\Big) = b. \tag{38}$$

Table 7: Simulated and approximated significance level when the signal $\{\boldsymbol{x}_\ell\}$ follows a VAR(1) model.

	$p = 2$		$p = 9$		$p = 36$	
b	Simulated	Approximated	Simulated	Approximated	Simulated	Approximated
3.5	0.097	0.097	0.065	0.057	0.036	0.042
4	0.063	0.068	0.036	0.030	0.013	0.019
4.5	0.038	0.047	0.018	0.019	0.006	0.008
5	0.033	0.032	0.011	0.012	0.003	0.003
5.5	0.022	0.021	0.005	0.007	0.002	0.001
6	0.015	0.014	0.003	0.004	0.0004	0.0005
6.5	0.006	0.009	0.002	0.002	0.0002	0.0002

Note that the solution of (38) can be obtained by a simple grid search when the dimension of θ is not too large.

The main proof technique for Theorem 3.3.1 is *change-of-measure*, which evaluates the boundary hitting probability of Gaussian processes [55, 76]. See Appendix B.3 for the derivation of (34) and Appendix B.4 for the proof of Theorem 3.3.1, when the dimension of parameter θ is 1 (i.e., $d = 1$). The proof can be generalized to multi-dimensional cases.

Although the theorem is an asymptotic result for large b, we find that this holds even for not very large b values in numerical studies. We verify the accuracy of Theorem 3.3.1 by comparing the approximated significance levels with simulated ones. In the experiment, we assume that the signal $\{\boldsymbol{x}_\ell\}$ follows a VAR(1) model, $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta\boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$, where $\theta \in \mathbb{R}$. Hence, $\boldsymbol{V}_\tau(\theta)$ has the form in (20). We further assume that the spatial correlation of the signal follows a spherical model, as defined in (17), with parameter $\rho = 0.3$. Set $N = 50$. The search space of θ is $\{0.1, 0.2, \cdots, 0.9\}$. In addition, the covariance matrix of the noise process $\boldsymbol{\Sigma}$ is assumed to be a p-by-p identity matrix. Simulation results are based on 5000 independent replications. Both simulated and approximated false alarm rates are reported in Table 7. As one can observe, the approximation is quite accurate.

In the proof of Theorem 3.3.1, we approximate the detection statistic $W(\tau, \theta)$ as a two-dimensional Gaussian random field. In the following, we verify that such an approximation is reasonable by simulation. We generate data under the null hypothesis, and verify the distribution of the detection statistic W for a set of fixed values of θ and τ. For the signal, we use a VAR(1) model, $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta\boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$ as the temporal correlation model and

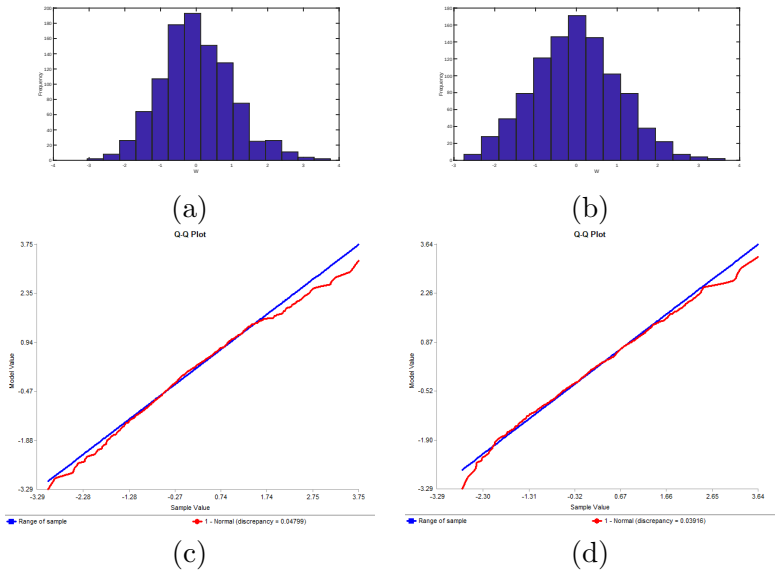


Figure 11: Histograms and q-q plots of $W(\theta, \tau)$ for fixed values of θ and τ: $\tau = 30$, $\theta = 0.3$ for (a) and (c); $\tau = 40$, $\theta = 0.2$ for (b) and (d).

a spherical model for spatial correlation model. We assume that the noise has the same spatial correlation structure as the signal. We set $N = 50$ and $p = 9$. Figure 11 shows the histograms and q-q plots of W for fixed values of θ and τ based on 1000 repetitions, which indicate that the Gaussian random field approximation is reasonable.

3.3.2 In-control Average Run Length ($\mathbf{ARL_0}$) for online $\mathsf{S^3T}$ statistic

In the online setting, the false alarm rate is characterized by the in-control average-run-length, which is equal to the expected stopping time of the procedure when there is no signal, denoted as $E_{\mathsf{H}_0}(\mathcal{T})$. The following theorem provides an approximation for $E_{\mathsf{H}_0}(\mathcal{T})$.

Theorem 3.3.2 (Approximation of ARL_0)**.** *Assume that $b \to \infty$. For the stopping time defined in* (30)*,*

$$E_{H_0}(\mathcal{T}) = (2\pi)^{\frac{d}{2}} \left(\int_{\theta \in \Theta} \frac{[b\xi_0(\omega,\theta)]^{\frac{d}{2}}}{\xi_0(\omega,\theta)} g(\omega,\theta) |H(\omega,\theta)|^{\frac{1}{2}} \frac{b^2 \mu(\omega,\theta)}{2\omega} \nu\Big(\sqrt{\frac{b^2\mu(\omega,\theta)}{\omega}}\Big) d\theta \right)^{-1} \Big(1 + o(1)\Big). \tag{39}$$

The derivation of Theorem 3.3.2 uses a similar technique based on the change-of-measure

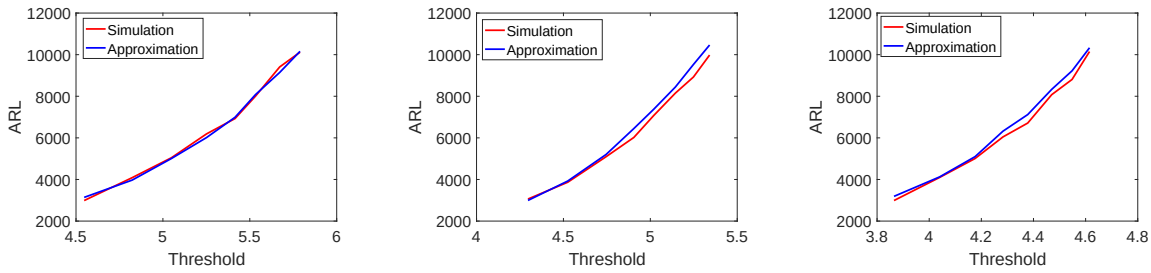


Figure 12: Comparison of approximated and simulated ARL for (a) $p = 1$, (b) $p = 2$, and (c) $p = 9$.

as in the derivation of Theorem 3.3.1. By Theorem 3.3.1, we can first obtain an approximation for the probability $\mathbb{P}_{\mathsf{H}_0}(\mathcal{T} \leq m)$, where $m > 0$ is fixed and sufficiently large:

$$\begin{aligned}\mathbb{P}_{\mathsf{H}_0}\Big(\mathcal{T} \leq m\Big) &= \mathbb{P}_{\mathsf{H}_0}\Big(\max_{\substack{\theta\in\Theta \\ 1\leq t\leq m}} W_t(\omega,\theta) \geq b\Big) \\ &= (2\pi)^{-\frac{d}{2}}\left(\sum_{t=1}^{m}\int_{\theta\in\Theta}\frac{[b\xi_0(\omega,\theta)]^{\frac{d}{2}}}{\xi_0(\omega,\theta)}g(\omega,\theta)|H(\omega,\theta)|^{\frac{1}{2}}\frac{b^2\mu(\omega,\theta)}{2\omega}\nu\Big(\sqrt{\frac{b^2\mu(\omega,\theta)}{\omega}}\Big)d\theta\right) + o(1).\end{aligned} \tag{40}$$

As argued in [56] and [57], the stopping time $\mathcal{T}$ is asymptotically exponentially distributed and is uniformly integrable. Hence, for large m, $\mathbb{P}_{\mathsf{H}_0}(\mathcal{T} \leq m) - [1 - \exp(-\lambda m)] \to 0$, where λ is approximately equal to the right hand side of (40) divided by m. Then by the first order Taylor expansion of an exponential term, we can obtain $E_{\mathsf{H}_0}(\mathcal{T}) \approx \lambda^{-1}$, which leads to (39).

The accuracy of Theorem 3.3.2 is verified by comparing the simulated and the approximated $E_{\mathsf{H}_0}(\mathcal{T})$. In the experiments, the signal $\{\boldsymbol{x}_\ell\}$ is generated by a VAR(1) model, $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta\boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$, where $\theta \in \mathbb{R}$. Hence, $\boldsymbol{V}_\tau(\theta)$ has the form in (20). Meanwhile, we assume that the spatial correlation of the signal follows a spherical model, as defined in (17), with parameter $\rho = 0.3$. The search space of parameter θ is $\{0.1, 0.2, \cdots, 0.9\}$. In addition, the covariance matrix of the noise process $\boldsymbol{\Sigma}$ is assumed to be a p-by-p identity matrix. The results based on 5000 replications are presented in Figure 12. The comparison between the simulated and approximated ARLs shows that the approximation in Theorem 3.3.2 is quite accurate.

3.4 Numerical Examples

In this section, we demonstrate the performance of the proposed detection procedures. Online change-point detection is the focus here, since it is the most relevant setting for our targeted applications of water quality monitoring. The performance comparison for offline change-point detection will be similar. We adopt the commonly used performance metric for sequential change detection, the expected detection delay (EDD) after a change has occurred. There is a tradeoff between the in-control average-run-length (ARL_0) and the EDD. Typically, we choose the threshold for each procedure so that its ARL_0 meets a pre-specified large value (e.g., 5000 or 10000), and hence there is rarely a false alarm. We also compare with other methods on simulated and real data.

3.4.1 Simulation

The detection procedure defined in (30) is compared with two other procedures: (i) an online detection procedure defined similarly to (30) using the quadratic score statistic $\widetilde{S}(\tau, \theta)$, and (ii) a multivariate cumulative sum (MCUSUM) procedure [19]. In the MCUSUM procedure, at each time step, a T^2 statistic [21] is calculated, which is combined with a CUSUM procedure.

In the experiment, the signal is generated from a VAR(1) model, $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta \boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$, with $p = 2$ and parameter $\theta = 0.5$. The spatial model of the signal follows the spherical model defined in (17) with $\rho = 0.3$. For both procedures, based on $\mathsf{S}^3\mathsf{T}$ and the quadratic score statistic, respectively, we use a window length $\omega = 50$ and the search space for the parameter θ, $\{0.1, 0.2, \cdots, 0.9\}$. Thresholds for all three procedures are calibrated so that they have the same false alarm rate $E_{\mathsf{H}_0}(\mathcal{T}) = 100$. To evaluate the expected detection delay, we assume the change occurs at $t = 1$. The mean of the signal $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}_\ell] = \mu \mathbb{1}_p$, $\mu \geq 0$. We explore different values of μ for the mean shift and γ for the magnitude of covariance matrix of the signal. If $\mu = 0$ and $\gamma > 0$, there is only a change in covariance; if both μ and γ are positive, then there are both mean shift and covariance change. Hence, the experiments demonstrate that the proposed detection procedure is suitable for both cases where there is either mean and/or covariance change.

Table 8: Simulated expected detection delay.

	$\mathsf{S}^3\mathsf{T}$					Quadratic score statistic					MCUSUM				
$\gamma\backslash\mu$	0	0.1	0.5	1	2	0	0.1	0.5	1	2	0	0.1	0.5	1	2
0.01	**97.27**	**59.08**	**6.37**	2.80	**1.49**	98.05	65.82	6.45	**2.77**	1.51	98.37	77.67	9.43	3.56	1.79
0.05	96.28	**57.96**	**5.95**	**2.72**	**1.49**	**95.32**	63.19	6.74	2.81	1.52	96.79	71.97	9.28	3.54	1.79
0.1	**72.93**	**53.16**	**6.04**	**2.78**	1.50	82.49	56.78	6.74	2.86	**1.49**	80.70	65.16	9.21	3.54	1.78
0.2	**65.32**	**46.16**	**5.96**	**2.77**	1.50	74.87	48.83	6.28	2.78	**1.47**	67.33	55.17	9.02	3.52	1.79
0.5	39.40	**30.32**	**5.81**	**2.78**	1.56	**37.07**	33.42	6.07	2.80	**1.50**	41.52	35.87	8.36	3.47	1.78
1	**20.91**	**19.42**	**5.65**	**2.75**	**1.51**	22.75	20.51	5.64	2.76	1.55	23.71	21.31	7.45	3.45	1.77

Table 8 reports the simulated EDD of three procedures based on 5000 repetitions. The smallest EDD values for each setting are marked in bold. The comparison shows that the two score statistic procedures, which capture both spatial and temporal correlation, outperform the MCUSUM procedure (which only captures the spatial correlation information). Such an advantage is more significant when the signal is weak, i.e., when γ or μ are both small. This demonstrates that incorporating temporal correlation information indeed improves detection performance. We also find that $\mathsf{S}^3\mathsf{T}$ outperforms the quadratic score statistic in many settings. This can be explained by that the quadratic score statistic needs to search more unknown parameters (the unknown μ), thus the statistic is noisier than $\mathsf{S}^3\mathsf{T}$ when there is no change. Therefore, to achieve the same ARL_0, the threshold for quadratic score statistic tends to be higher, which may cause a larger detection delay. Given that $\mathsf{S}^3\mathsf{T}$ enjoys tractable theoretical analysis and an accurate approximation for its false alarm rate, it is a good option for practitioners.

3.4.2 Real data example: Solar flare detection

We apply our detection procedure to a real dataset, which is acquired by the Solar Data Observatory [37]. The data is a video sequence that contains an abrupt emergence of a solar flare that occurs around time $t = 227$. In this video, the normal state is a sequence of slowly drifting image of the solar surface, and the changes are much brighter transient solar flares. Figure 13 shows a snapshot when a solar flare occurs at $t = 227$.

The size of the images is 232×292 pixels. After vectoring the images, this leads to $67,744$ dimensional vectors. Due to the high dimensionality, it is computationally expensive to apply our detection procedure on the original images directly. Hence, we apply the

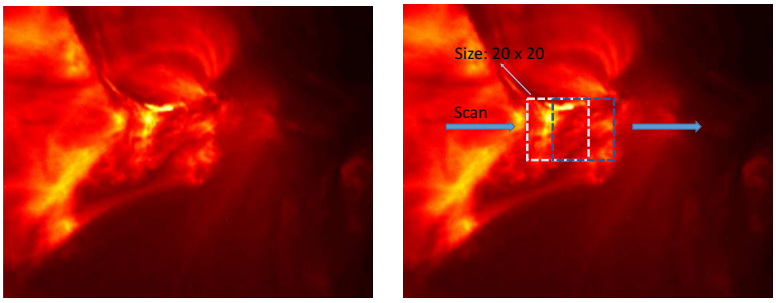


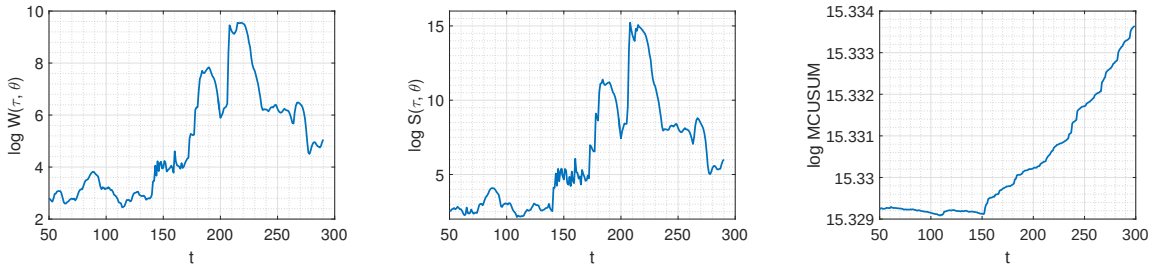
Figure 13: Detection of solar flare at $t = 227$: (left) snapshot of the original SDO data at $t = 227$; (right) overlapping image patches for dimensionality reduction.

spatial scanning scheme proposed in the previous chapter. We break the original image into overlapping patches of dimension 20×20, as demonstrated in the right panel of Figure 13. The detection statistic is calculated for each image patch (of dimension $p = 400$). Then, we take the maximum of the detection statistic over all patches.

We assume that before the solar flare, the data form a white noise process with no spatial and temporal correlation. The mean and variance of the noise process are estimated by the first 50 samples in the sequence. For the signal, we use a VAR(1) model, $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta \boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$ to capture the temporal correlation. The spatial model of the signal is captured by a spherical model defined in (17). Online procedures are implemented with window length $\omega = 10$. Figure 14(a), 14(b) and 14(c) show the values of $\mathsf{S}^3\mathsf{T}$ statistic, the quadratic score statistic and MCUSUM statistic on a logarithmic scale, respectively. Since in this case, we do not have the ground truth, we cannot evaluate the true EDD. However, as we can observe, both $\mathsf{S}^3\mathsf{T}$ and the quadratic score statistics obtain peak detection statistics at around $t = 227$, and another solar flare at around $t = 173$, indicating both statistics can successfully detect the emergence of solar flares. However, MCUSUM statistic misses both solar flares.

3.4.3 Case study: Water quality monitoring

In this section, we consider a case study of real-time water quality monitoring for the Altamaha River network based on data generated by the SWMM model. The goal is to detect contaminant spills that pollute the river as quickly as possible. Backgrounds on the



(a) $\mathsf{S}^3\mathsf{T}$ (b) Quadratic score statistic (c) MCUSUM

Figure 14: Detection statistics on logarithmic scale.

Altamaha River and SWMM model are presented in Section 2.5.1.

In the case study, the tail-up model (discussed in Section 2.5.2) with an exponential correlation function is adopted as the spatial model for the data. We assume that both the signal and the noise share the same spatial correlation structure. For temporal correlation, we use a VAR(1) model $\boldsymbol{x}_\ell = (1-\theta)\boldsymbol{\mu} + \theta\boldsymbol{x}_{\ell-1} + \boldsymbol{\epsilon}_\ell$ to capture the temporal correlation of a contaminant spill as suggested in [8] and [9]. We apply the online change-point detection procedure based on $\mathsf{S}^3\mathsf{T}$ to detect contaminant spills in the Altamaha River network. We also compare it with two other methods: (i) online detection based on the quadratic score statistic, and (ii) the Hotelling's T^2 chart. Among the 100 nodes on the river network, 10 of them (nodes 1, 15, 19, 33, 36, 50, 58, 67, 84, 95, marked by red stars in Figure 8(c)) are used as possible contaminant spill locations, and the rest 90 nodes are used for collecting measurements every 15 minutes. In each replication, we run SWMM to simulate the river network during a 10-day period. A single instantaneous spill is generated, with a spill location randomly selected from the ten possible locations. The spill starting time is uniformly distributed between the first 15 to 20 hours. The intensity of the contaminant spills follows a uniform distribution, and we consider three different levels: U(10, 100) (low), U(100, 250) (medium), and U(250, 500) (high) in units of gram/liter.

The thresholds for the three detection procedures are adjusted so that the in-control ARL_0s are 10 days (960 samples). For the two procedures based on $\mathsf{S}^3\mathsf{T}$ and the quadratic score statistic, the length of the sliding window is chosen as 12.5 hours (50 samples). Table 9 reports the average and standard error of detection delays obtained from 100 simulated spills. For spills with high intensity, all three methods achieve similar performance regarding

Table 9: Simulated expected detection delay in hours (numbers in parentheses are standard errors).

Spill Intensity	$\mathsf{S}^3\mathsf{T}$	Quadratic Score Statistic	T^2
low	**38.285 (3.655)**	45.822 (4.675)	52.959 (5.035)
medium	**26.301 (1.679)**	28.522 (1.873)	30.753 (2.192)
high	25.519 (1.697)	**25.489 (1.667)**	25.563 (1.860)

detection delay, as strong signals are easier to be detected. However, when the signal is relatively weak (low and medium spill intensity), the proposed detection statistic $\mathsf{S}^3\mathsf{T}$ significantly outperforms the other two methods.

3.5 Conclusions

In this chapter, we propose a novel efficient score statistic $\mathsf{S}^3\mathsf{T}$ to detect the emergence of a spatial-temporal signal from a noisy background in both the offline and online settings. The statistic captures the spatial and temporal correlation simultaneously and enjoys a relatively low computational cost. An accurate approximation for its false alarm rate is presented. Numerical results based on simulated data, real solar flare data, and a case study of water quality monitoring show that the proposed $\mathsf{S}^3\mathsf{T}$ statistic has a clear advantage than existing methods.

CHAPTER IV

COMBINING CONSTRAINED BAYESIAN OPTIMIZATION AND SPATIO-TEMPORAL CHANGE-POINT DETECTION FOR SENSOR NETWORK DESIGN

The advances of sensor technology have enabled online monitoring for complicated systems. A sensor network consists of a group of sensors dispersed in space and collects multiple data streams in real time. Design of a sensor network answers not only where the sensors should be placed in the space but also how the data streams collected by the sensors should be processed and analyzed. In this chapter, we focus on designing a sensor network for water quality monitoring on a river system. The goal is to generate timely information regarding water quality and enable quick detections on undesired contamination events.

In practice, sensors are usually subject to random measurement error. [26] point out that decision making based on a sensor network is unreliable if the inaccuracy of data is not handled by a scientific approach. In water quality monitoring, the detection of contamination based on sensor data is a statistical problem. [26] adopt statistical process control (SPC) methods including a Shewhart chart and a CUSUM chart in sensor network design when measurement error exists. However, these approaches do not capture the spatial and temporal correlation in the sensor data. In Chapter 3, we propose the score statistic $\mathsf{S}^3\mathsf{T}$, which detects the emergence of a signal from noisy background. The $\mathsf{S}^3\mathsf{T}$ statistic captures both spatial and temporal correlation in the data and hence is particularly good at detecting weak signals.

In this chapter, (i) we formulate the problem of sensor network design for water quality monitoring on river systems as a joint problem of constrained black-box function optimization and online statistical change-point detection; (ii) we propose a new algorithm called Confidence-Set based Constrained Bayesian Optimization (CSCBO), which provides a flexible framework to handle noisy black-box function constraints and is easy to implement;

(iii) we extend the algorithm to tackle with a challenge that arises specifically in the sensor network design problem: we adopt the Wasserstein similarity metric to enable CSCBO to solve problems with high-dimensional binary search space; and (iv) finally we combine CSCBO with the $\mathsf{S}^3\mathsf{T}$ statistic to find robust sensor networks for the Altamaha river and show advantages of the proposed methods.

The rest of the chapter is organized as follows. Section 4.1 provides backgrounds on the constrained black-box function optimization problem and Bayesian optimization approach. Section 4.2 presents the proposed CSCBO algorithm. Section 4.3 presents the formulation and methodologies for sensor network design on river systems. Section 4.4 contains experimental results. Section 4.5 concludes the chapter.

4.1 Background

In this section, we provide general settings of the constrained black-box function optimization problems and a brief overview of the BO method.

4.1.1 Constrained black-box function optimization

Suppose a stochastic system of interest have $J+1$ performance metrics. One is considered as the primary performance metric and the rest are used as guardrail performance metrics. We seek to find the optimal input parameter of the system that leads to the best primary performance metric subject to nonnegative constraints on each guardrail performance metric. In mathematical notation, we consider the following constrained optimization problem,

$$\begin{aligned} \max_{\boldsymbol{x}\in\mathfrak{X}} \quad & f_0(\boldsymbol{x}) \\ \text{s.t.} \quad & f_j(\boldsymbol{x}) \geq 0, \quad \forall j \in \{1, \cdots, J\} \end{aligned} \tag{41}$$

where $\boldsymbol{x}$ is a D-dimensional vector of input variables, $\mathfrak{X} \subset \mathbb{R}^D$ is a bounded search space, $f_0 : \mathfrak{X} \to \mathbb{R}$ is the primary performance function and $f_j : \mathfrak{X} \to \mathbb{R}$, $(j \geq 1)$ denote the jth guardrail performance function. Note that the search space $\mathfrak{X}$ is assumed to have been discretized if the original domain of the functions is continuous, and hence, $\mathfrak{X}$ is a discrete set and contains a finite number of points.

The functions f_j, $j = 0, \cdots, J$ are considered to be unknown black-box functions that are expensive to evaluate and subject to noise. In other words, for any point $\boldsymbol{x} \in \mathfrak{X}$, $f_j(\boldsymbol{x})$

cannot be evaluated analytically. Usually $f_j(\boldsymbol{x})$, $j = 0, \cdots, J$ are expectations of random estimates based on simulations for the performance metrics of the system given a parameter $\boldsymbol{x}$,

$$f_j(\boldsymbol{x}) = \mathrm{E}[y_j(\boldsymbol{x})], j \in \{0, \cdots, J\}.$$

Here $y_j(\boldsymbol{x})$ is a simulation estimate of the jth performance metric, which we can be a single observation or a sample average of multiple observations. We assume that

$$y_j(\boldsymbol{x}) \sim \mathbb{N}\Big(f_j(\boldsymbol{x}), \lambda_j(\boldsymbol{x})\Big),$$

where $\lambda_j(\cdot)$ is referred to as the sampling variance of the jth performance metric. In practice, $\lambda_j(\cdot)$ is not necessarily known and should be estimated if unknown. From one run of simulation with input parameter $\boldsymbol{x}$, we are able to evaluate all performance metrics jointly and obtain a vector of observations $[y_0(\boldsymbol{x}), y_1(\boldsymbol{x}), \cdots, y_J(\boldsymbol{x})]$.

4.1.2 Bayesian Optimization

Bayesian optimization (BO) is originally proposed to solve the unconstrained version of problem (41), i.e., $J = 0$. A BO algorithm is a sequential procedure and typically consists of two components: a surrogate model characterizing the function and an acquisition function guiding evaluations. In practice, Gaussian Process (GP) is most widely adopted to model the black-box objective function due to its flexibility and tractability. In the initial stage of the optimization, a GP prior is put over the function which is specified by a mean function $\mu_0(\boldsymbol{x}) : \mathfrak{X} \to \mathbb{R}$ and a kernel function $K_0(\boldsymbol{x}, \boldsymbol{x}') : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$,

$$f_0(\boldsymbol{x}) \sim GP\Big(\mu_0(\boldsymbol{x}), K_0(\boldsymbol{x}, \boldsymbol{x}')\Big).$$

Suppose that we have evaluated f_0 at n points $\boldsymbol{x}^{(1:n)} := \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(n)}\}$, and obtained corresponding observations $y_0^{(1:n)} := \{y_0^{(1)}, y_0^{(2)}, \cdots, y_0^{(n)}\}$. Then, we can calculate the posterior distribution on f_0 by combining the prior and the observations based on the Bayes rule. The posterior is still a GP

$$f_0(\boldsymbol{x})\Big|\boldsymbol{x}^{(1:n)}, y_0^{(1:n)} \sim GP\Big(\mu_0^{(n)}(\boldsymbol{x}), K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x}')\Big),$$

where

$$\mu_0^{(n)}(\boldsymbol{x}) = \mu_0(\boldsymbol{x}) + K_0(\boldsymbol{x}, \boldsymbol{x}^{(1:n)})\Big(K_0(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}) + \text{diag}\{\lambda(\boldsymbol{x}^{(1)}), \cdots, \lambda(\boldsymbol{x}^{(n)})\}\Big)^{-1}\big(y_0^{(1:n)} - \mu_0(\boldsymbol{x}^{(1:n)})\big), \tag{42}$$

and

$$K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x}') = K_0(\boldsymbol{x}, \boldsymbol{x}') - K_0(\boldsymbol{x}, \boldsymbol{x}^{(1:n)})\Big(K_0(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}) + \text{diag}\{\lambda(\boldsymbol{x}^{(1)}), \cdots, \lambda(\boldsymbol{x}^{(n)})\}\Big)^{-1} K_0(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}'). \tag{43}$$

Typically, a BO algorithm determines the next evaluation point by maximizing an acquisition function which relies on the posterior distribution of f_0,

$$\boldsymbol{x}^{(n+1)} = \arg\max_{\boldsymbol{x} \in \mathfrak{X}} \ acquisition\_function(\boldsymbol{x}).$$

Expected improvement (EI) [23] is one of the most commonly used acquisition function, which is calculated as follows,

$$\begin{aligned} \text{EI}(\boldsymbol{x}|f_0^\star) &= \text{E}[\max(0, f_0(\boldsymbol{x}) - f_0^\star)] \\ &= (\mu_0^{(n)}(\boldsymbol{x}) - f_0^\star)\Phi(Z) + \sqrt{K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x})}\phi(Z), \end{aligned} \tag{44}$$

where $Z = \frac{\mu_0^{(n)}(\boldsymbol{x}) - f_0^\star}{\sqrt{K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x})}}$, $f_0^\star$ is the current best function value found and the expectation is taken over the posterior distribution of f_0. Note that the original EI proposed in [23] assumes that $\lambda(\boldsymbol{x}) = 0, \ \forall \boldsymbol{x} \in \mathfrak{X}$, i.e., the function to be optimized is not subject to evaluation noise. When noise exists, calculating (44) is difficult as $f_0^\star$ is unknown. [2] propose to modify EI by replacing $f_0^\star$ in (44) with the GP posterior mean estimate of the best function value $\mu_0^{(n)^\star} = \max_{\boldsymbol{x} \in \mathfrak{X}} \mu_0^{(n)}(\boldsymbol{x})$,

$$\text{EI}(\boldsymbol{x}|\mu_0^{(n)^\star}) = (\mu_0^{(n)}(\boldsymbol{x}) - \mu_0^{(n)^\star})\Phi(Z) + \sqrt{K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x})}\phi(Z). \tag{45}$$

For the rest of the chapter, we use EI to refer to the modified version defined in (45).

Another example of acquisition function is the upper confidence bound (UCB) [62]. UCB returns the next evaluation point with the highest upper confidence interval based on the posterior distribution of f_0,

$$\text{UCB}(\boldsymbol{x}, \alpha) = \mu_0^{(n)}(\boldsymbol{x}) + z_\alpha \sqrt{K_0^{(n)}(\boldsymbol{x}, \boldsymbol{x})}, \tag{46}$$

where z_α is the αth quantile of the standard normal distribution.

While EI and UCB are powerful methods and are successful in many of applications, they deal with unconstrained problems only. In the next section, we propose an efficient and practical algorithm that is capable of handling optimization problem with noisy black-box function constraints.

4.2 Confidence-Set based Constrained Bayesian Optimization

In this section, we present a new algorithm for problem (41) called Confidence-Set based Constrained Bayesian Optimization (CSCBO).

To solve the problem defined in (41), CSCBO puts independent GP priors over each function f_j at the initialization step,

$$f_j(\boldsymbol{x}) \sim GP\Big(\mu_j(\boldsymbol{x}), K_j(\boldsymbol{x}, \boldsymbol{x}^{'})\Big), \quad j = 0, \cdots, J, \tag{47}$$

where $\mu_j()$ and $K_j()$ are the prior mean and prior kernel functions for f_j. In the nth iteration, suppose the algorithm chooses to evaluate the functions at $\boldsymbol{x}^{(n)}$ and obtains a vector of observations $[y_0^{(n)}, y_1^{(n)}, \cdots, y_J^{(n)}]$. The algorithm first updates the posterior distribution for each function independently,

$$f_j(\boldsymbol{x}) \sim GP\Big(\mu_j^{(n)}(\boldsymbol{x}), K_j^{(n)}(\boldsymbol{x}, \boldsymbol{x}^{'})\Big), \quad j = 0, \cdots, J, \tag{48}$$

where $\mu_j^{(n)}()$ and $K_j^{(n)}()$ denote the posterior mean and kernel functions for f_j after n evaluations and can be calculated using (42) and (43).

Next, we construct a *confidence set* (denoted as CS_1) using the posterior distributions, which is a subset of the search space $\mathfrak{X}$ and eliminates points based on the criteria described in the following. We first define a reward function for each constraint as follows,

$$r_j(\boldsymbol{x}) = \begin{cases} 1, & \text{if } f_j(\boldsymbol{x}) \geq 0; \\ 0, & \text{otherwise,} \end{cases}$$

where $j = 1, \cdots, J$. We then calculate the expected reward function with respect to the posterior distributions of the constraint functions,

$$\mathrm{E}\Big[r_j(\boldsymbol{x})\Big] = \mathrm{E}\Big[\mathbb{1}\Big\{f_j(\boldsymbol{x}) \geq 0\Big\}\Big] = \mathbb{P}\Big(f_j(\boldsymbol{x}) \geq 0\Big) = \Phi\Big(\frac{\mu_j^{(n)}(\boldsymbol{x})}{\sqrt{K_j^{(n)}(\boldsymbol{x}, \boldsymbol{x})}}\Big),$$

Algorithm 1 CSCBO

Input: GP priors $\mu_j()$, $K_j()$, $j = 0, \cdots, J$; h_1; h_2.
Initialization: set $n = 0$, CS_1=$\mathfrak{X}$, and randomly pick $\boldsymbol{x}^{(0)} \in CS_1$.
while stopping criteria is not met **do**
  evaluate at $\boldsymbol{x}^{(n)}$ and obtain $[y_0^{(n)}, y_1^{(n)}, \cdots, y_J^{(n)}]$.
  update GP posteriors $\mu_j^{(n)}()$ and $K_j^{(n)}()$ for $j = 0, \cdots, J$.
  set $CS_1 = \Big\{\boldsymbol{x} \in \mathfrak{X} : \Phi\Big(\frac{\mu_j^{(n)}(\boldsymbol{x})}{\sqrt{K_j^{(n)}(\boldsymbol{x},\boldsymbol{x})}}\Big) \geq h_1 \ \forall\ j \in \{1, \cdots, J\}\Big\}$.
  obtain $\mu_0^{(n)\star} = \max_{\boldsymbol{x} \in CS_1}\ \mu_0^{(n)}(\boldsymbol{x})$ and $\boldsymbol{x}^{(n+1)} = \arg\max_{\boldsymbol{x} \in CS_1}\ \mathrm{EI}(\boldsymbol{x}|\mu_0^{(n)\star})$.
  set $CS_2 = \Big\{\boldsymbol{x} \in \mathfrak{X} : \Phi\Big(\frac{\mu_j^{(n)}(\boldsymbol{x})}{\sqrt{K_j^{(n)}(\boldsymbol{x},\boldsymbol{x})}}\Big) \geq h_2 \ \forall\ j \in \{1, \cdots, J\}\Big\}$.
  report current optimal feasible solution $\boldsymbol{x}_\star^{(n)} = \arg\max_{\boldsymbol{x} \in \mathrm{CS}_2} \mu_0^{(n)}(\boldsymbol{x})$.
  $n = n + 1$.
end while
return optimal feasible solution $\boldsymbol{x}_\star^{(n)}$.

where $\mathbb{1}\{\cdot\}$ denotes an indicator function and $\Phi(\cdot)$ is the cumulative density function (CDF) of a standard normal distribution. Note that $f_j(\boldsymbol{x})$, $j = 1, \cdots, J$ are treated as Gaussian random variables with mean $\mu_j^{(n)}(\boldsymbol{x})$ and variance $K_j^{(n)}(\boldsymbol{x}, \boldsymbol{x})$. CS_1 is constructed based on the expected rewards for all constraints,

$$CS_1 = \{\boldsymbol{x} \in \mathfrak{X} : \mathrm{E}[r_j(\boldsymbol{x})] \geq h_1,\ \forall j = 1, \cdots, J\}, \tag{49}$$

where $h_1 \in [0, 1]$ is a pre-specified constant. Namely, CS_1 contains the points whose probabilities of satisfying each constraint are at least h_1. In each iteration, the algorithm selects the next evaluation point from CS_1 based on the modified EI criterion defined in (45),

$$\boldsymbol{x}^{(n+1)} = \underset{\boldsymbol{x} \in CS_1}{\arg\max}\ \mathrm{EI}(\boldsymbol{x}|\mu_0^{(n)\star}). \tag{50}$$

For reporting optimal feasible solution found so far, another confidence set CS_2 is constructed with a different parameter h_2. The point with the largest posterior mean in CS_2 is reported as the best feasible solution. The formal procedure of CSCBO is presented in Algorithm 1.

A tradeoff that one faces when dealing with a constrained black-box function optimization problem is whether to search for a promising solution in terms of the primary performance metric first and then check the feasibility or attempt to identify the boundary between feasible and infeasible regions first and then search for promising solutions. While

eventually an optimal feasible solution is desired, the two manners may lead to different search schemes. We design CSCBO based on the idea that improving the primary performance metric should be of higher priority. Specifically, the feasibility of a solution matters only if the solution is likely to have a large value of the primary performance metric. Thus, in CSCBO, the acqusition function itself is purely based on the primary performance function f_0, and the constraints only affect the search procedure via the confidence set CS_1. CS_1 restricts the region where the next evaluation point is chosen and is supposed to only eliminate the points with strong evidence to be infeasible. Thus, h_1 should be a relatively small value. CS_1 is initialized to include all points in $\mathfrak{X}$ and hence the restriction on the search space is small in the early iterations of the optimization. In practice, computation budget is not unlimited and the optimization algorithm is subject to termination within finite number of iterations. The confidence set CS_2 is constructed for the purpose of reporting an optimal feasible solution after a finite number of iterations. The optimal feasible solution identified in the nth iteration, which is denoted as $\boldsymbol{x}_\star^{(n)}$, is the point with the highest posterior mean in CS_2. To gain more confidence in the feasibility of $\boldsymbol{x}_\star^{(n)}$, we recommend that h_2 is set as a relatively large value, and hence there is strong evidence for points in CS_2 to be feasible. A simulation study for different choices of h_1 and h_2 is presented in Section 4.2.2.

Stopping criteria. One may choose to terminate the algorithm after a pre-specified maximum number of iterations. Alternatively, we can terminate the algorithm after the same point has been reported as the optimal feasible solution in k consecutive iterations. A recommended value is $k = 50$.

4.2.1 Connection to confidence bounds

An alternative way to construct a confidence set is based on confidence bounds of a GP. We define an α-level confidence bound of the posterior GP for function f_j after n evaluations as follows,

$$CB_j(\boldsymbol{x}, \alpha) = \mu_j^{(n)}(\boldsymbol{x}) + z_\alpha \sqrt{K_j^{(n)}(\boldsymbol{x}, \boldsymbol{x})},$$

where z_α is the αth upper quantile of a standard normal distribution, i.e., $\mathbb{P}(Z \geq z_\alpha) = \alpha$ and $Z \sim \mathbb{N}(0,1)$. Then, we can construct CS_1 as follows,

$$CS_1 = \Big\{ \boldsymbol{x} \in \mathfrak{X} : CB_j(\boldsymbol{x}, \alpha_1) \geq 0 \ \forall \ j \in \{1, \cdots, J\} \Big\}, \tag{51}$$

where $\alpha_1 \in [0,1]$ is a constant. In the following, we demonstrate that the two confidence sets defined in (49) and (51) are equivalent if $\alpha_1 = h_1$. First we have,

$$\mathbb{P}\Big(f_j(\boldsymbol{x}) \geq CB_j(\boldsymbol{x}, \alpha_1) \Big) = \mathbb{P}\Big(f_j(\boldsymbol{x}) \geq \mu_j^{(n)}(\boldsymbol{x}) + z_{\alpha_1} \sqrt{K_j^{(n)}(\boldsymbol{x}, \boldsymbol{x})} \Big) = \alpha_1.$$

Also, we have that $\mathbb{P}(f_j(\boldsymbol{x}) \geq t)$ is a monotone decreasing function of t as $f_j(x)$ is treated as a random variable. Therefore,

$$CB_j(\boldsymbol{x}, \alpha_1) \geq 0 \Leftrightarrow \mathrm{E}[r_j(\boldsymbol{x})] = \mathbb{P}\Big(f_j(\boldsymbol{x}) \geq 0 \Big) \geq \alpha_1 = h_1.$$

4.2.2 Choices for h_1 and h_2

Here we demonstrate the impact of h_1 and h_2 values on the performance of CSCBO. We solve the following problem,

$$\begin{aligned} \max_{\boldsymbol{x} \in [0,100]} \quad & \mathrm{E}\Big[\tfrac{\sin^6(0.05\pi x)}{2^{2(\frac{x-10}{80})^2}} + \epsilon_0 \Big] \\ \text{s.t.} \quad & \mathrm{E}\Big[\sin(0.1\pi x) + H + \epsilon_1 \Big] \geq 0, \end{aligned}$$

where H is a constant and ϵ_0 and ϵ_1 are independent normal random variables with mean 0 and variance σ_ϵ^2. We set $H \in [-0.8, -0.99]$ to create scenarios where the proportion of feasible solution is relatively small ($H = -0.8$) and extremely small ($H = -0.99$). σ_ϵ is set as 0 or 0.01 to test the performance of CSCBO on deterministic and stochastic functions, respectively. The search space is uniformly discretized into 1000 points. We define a utility function $u(\boldsymbol{x})$ which is equal to $f_0(\boldsymbol{x})$ if $\boldsymbol{x}$ is feasible and otherwise the worst objective function value achievable in the search space. In each iteration, we record the utility of the current optimal feasible solution reported by the algorithm $u(\boldsymbol{x}_\star^{(n)})$ and compute the utility gap $|u(\boldsymbol{x}_\star^{(n)}) - u(\boldsymbol{x}_\star)|$, where $\boldsymbol{x}_\star$ is the true optimal feasible solution of the problem. The algorithm is initialized with a random point in the search space.

Figure 15 shows the mean of the utility gap for 500 replications with different randomized initialization. We can observe that the algorithm performs better if h_1 is set to a

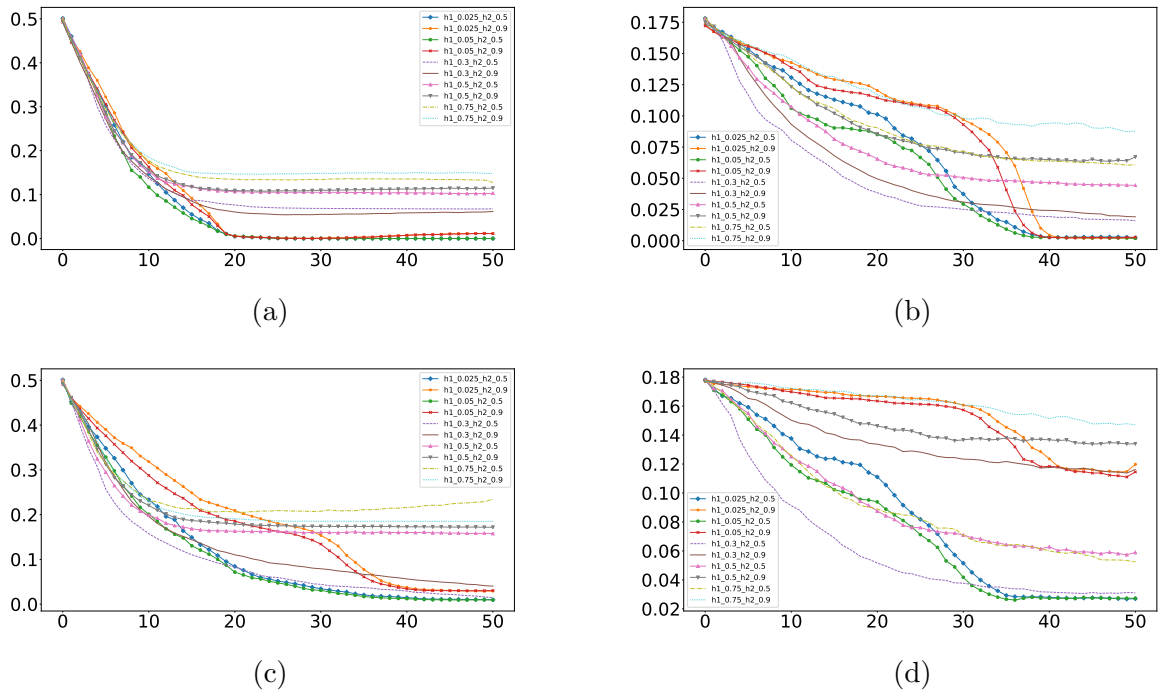


Figure 15: Mean utility gap between the true optimal feasible solution and the solution found by CSCBO: (a) deterministic, $H = -0.8$, (b) deterministic, $H = -0.99$, (c) stochastic, $H = -0.8$ and (d) stochastic, $H = -0.99$.

smaller value (e.g., $h_1 = 0.025$ or $h_1 = 0.05$), which achieves smaller mean utility gap when converging, than a larger value ($h_1 \in [0.3, 0.5, 0.75]$). The reason is that a large h_1 may restrict the search space too much in early iterations of the optimization when only a few function evaluations can be used to update the posterior distribution of the GPs, which leads to a lack of exploration. Therefore, h_1 should be set to a small value to avoid this issue. We also find that the algorithm converges quicker with $h_2 = 0.5$ than $h_2 = 0.9$. Such an advantage is more obvious for problems with lower proportion of feasible solutions ($H = -0.99$). This is due to that $h_2 = 0.9$ is sometimes too strict for the algorithm to find any feasible solution when the true proportion of feasible solutions is small. Based on these observations, we should avoid using a large h_2 value in practice.

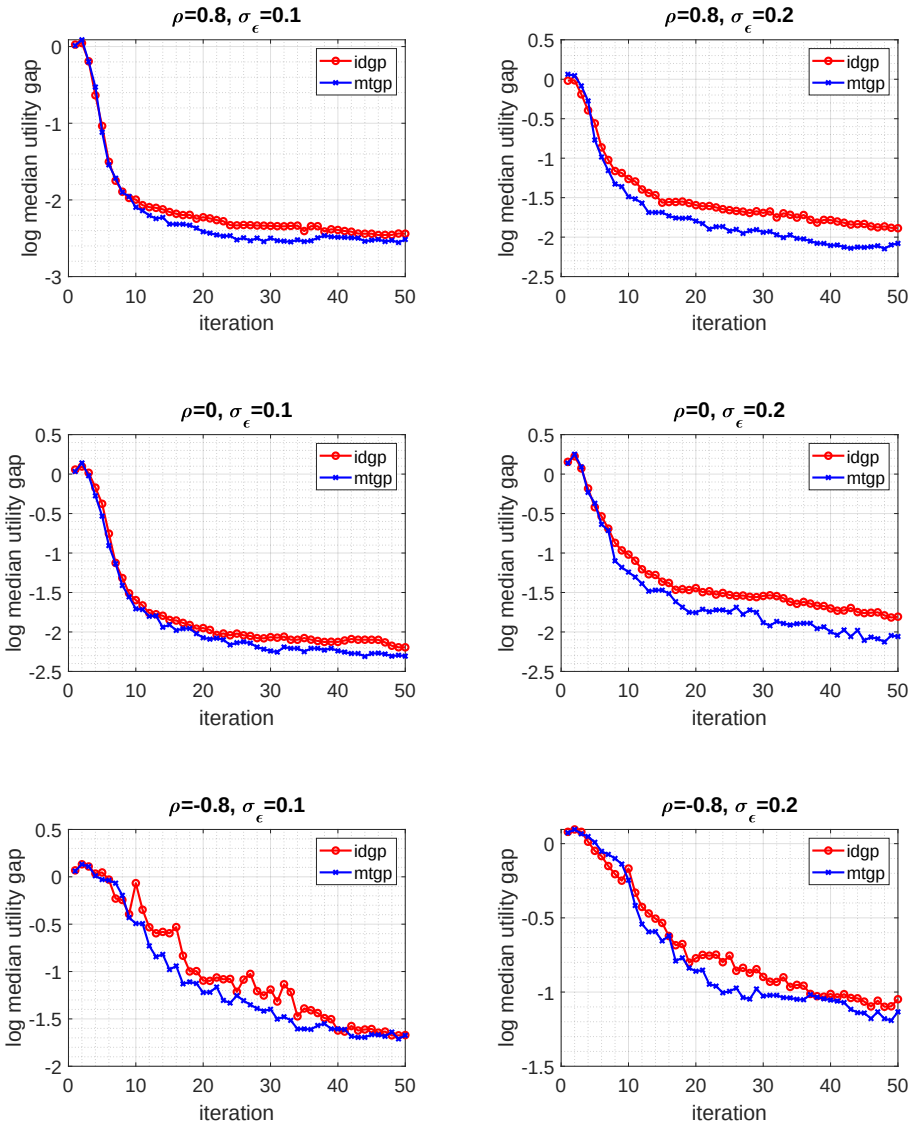


Figure 16: Comparison on log median utility gap between CSCBO with independent GPs (idgp) and multi-task GPs (mtgp).

4.2.3 Extend to multi-task Gaussian processes (MTGP)

[4] extends the Gaussian processes to the case of multi-output functions by defining a covariance function $K_{\text{multi}}((\boldsymbol{x}, t), (\boldsymbol{x}', t'))$ between functions,

$$K_{\text{multi}}((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = K_t(t, t') \otimes K_x(\boldsymbol{x}, \boldsymbol{x}'),$$

where K_x is the covariance function between input points and K_t is the covariance between functions. The posterior distribution of a multi-task GP can be obtained by the common approach once the covariance function K_{multi} is given. The benefit of MTGP is that information about multiple functions can be shared in the joint model. To take advantage of such benefit, we extend CSCBO by modeling all functions f_j, $j = 0, \cdots, J$ using a MTGP and jointly obtain the posterior mean and variances for each function in every iteration. The rest steps proceed in the same way.

In the following, we demonstrate the performance of CSCBO with MTGP using synthetic problems in the form of (41) with $J = 1$. The search space is 1000 uniformly discretized points in $[0, 1]$. In each problem, an objective function and a constraint are randomly generated by a zero-mean MTGP with a squared exponential kernel of unit amplitude and length scale $\ell = 0.1$. The correlation between the two GPs used for generating objective functions and constraints is set to -0.8, 0 and 0.8. The sampling standard deviation σ_ϵ is set to 0.1 and 0.2. Note that in the experiment, the true information of the GP models used for generating functions is not given to the algorithm. We define the utility function $u(\boldsymbol{x})$ in the same way as in Section 4.2.2. For each setting, we randomly generate 50 problems and repeat the solving process 10 times with random initialization for each problem. Figure 16 presents the median utility gap on logarithmic scale for CSCBO based on MTGP and independent GPs. We can observe that CSCBO with MTGP outperforms independent GPs in all cases, although the difference is small when the sample variance is small. This indicates that MTGP indeed helps to improve the efficiency of CSCBO by capturing the correlation between the objective function and constraint. Interestingly, the advantage of MTGP exists even when $\rho = 0$. The reason is that although independently, the objective function and the constraint are randomly generated by GPs with same hyperparameters and hence tends to behave similarly.

4.3 Combine CSCBO and $\mathsf{S}^3\mathsf{T}$ for Sensor Network Design

In this section, we combine CSCBO and $\mathsf{S}^3\mathsf{T}$ to solve the problem of sensor network design for river water quality monitoring. The goal of the network is to quickly and accurately

detect contaminant events on the river.

4.3.1 Formulation

We first formulate the problem of optimal sensor placement as a constrained black-box function optimization problem as defined in (41) under the assumption that sensors are error-free, i.e., sensor measurement error does not exist. A river network is assumed to have D nodes, indexed from 1 to D and each one is a potential location to place a sensor. Denote the number of sensors as M and assume $M < D$. The decision variable $\boldsymbol{x}$ is a set of nodes where the M sensors are placed. $\boldsymbol{x}$ is a D-dimensional binary vector with $\boldsymbol{x}_i = 1$ if a sensor is placed on the ith node and 0 otherwise. The search space $\mathfrak{X} = \{\boldsymbol{x} \in BV^D : |\boldsymbol{x}|_1 = M\}$, where BV^D denotes a D-dimensional binary vector space and $|\cdot|_1$ denotes the L_1 norm of a vector.

The sensors collect concentration data in real-time. At time t, the M sensors obtain a vector of concentration measurements $\boldsymbol{c}_t = [c_{1t}, \cdots, c_{Mt}]$, where c_{it} denotes the data measured by the ith sensor at time t. Under the error-free assumption, the concentration measurements are collected with 100% accuracy. A detection statistic W_t is constructed based on the concentration measurements collected at and prior to t. A simple example of a detection statistic is

$$W_t = \max(c_{1t}, \cdots, c_{Mt}). \tag{52}$$

The monitoring system will raise an alarm at time t warning a contaminant is detected if W_t exceeds an pre-specified threshold b. Since sensors are assumed to be error-free, if an alarm is raised, a true contaminant spill is detected successfully. Denote t_0 as the starting time of a contaminant spill event and $t_a(\boldsymbol{x})$ as the time stamp when the sensor network at $\boldsymbol{x}$ raises an alarm,

$$t_a(\boldsymbol{x}) = \inf\Big\{t : W_t \geq b\Big\}.$$

Then we define detection delay $T(\boldsymbol{x})$ as follows,

$$T(\boldsymbol{x}) = t_a(\boldsymbol{x}) - t_0. \tag{53}$$

$T(\boldsymbol{x})$ is the amount of time elapsed between the start and the detection of a contaminant event. Note that it is possible that a sensor network fails to detect the contaminant event.

In such a case, the sensor network will never raise an alarm and $T(\boldsymbol{x}) = \infty$. We define an indicator function

$$R(\boldsymbol{x}) = \begin{cases} 0, & \text{if the sensor network fails to detect a contaminant event, (i.e., } T(\boldsymbol{x}) = \infty); \\ 1, & \text{otherwise.} \end{cases} \tag{54}$$

Note that both $T(\boldsymbol{x})$ and $R(\boldsymbol{x})$ are random variables and can only be observed via stochastic simulations.

Adopting the same formulation in [41], we formulate the problem of optimal sensor placement for river water quality monitoring network with error-free sensors as follows,

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathfrak{X}} \quad & \mathrm{E}[T(\boldsymbol{x})\big|R(\boldsymbol{x}) = 1] \\ \text{s.t.} \quad & \mathrm{E}[R(\boldsymbol{x})] \geq q, \end{aligned} \tag{55}$$

where $0 \leq q \leq 1$. We refer to $\mathrm{E}[T(\boldsymbol{x})\big|R(\boldsymbol{x}) = 1]$ as the conditional expected detection delay and $\mathrm{E}[R(\boldsymbol{x})]$ as the reliability, both of which are stochastic black-box functions.

4.3.2 Measurement error

In practice, sensors are usually subject to measurement error, which means the concentration measurements collected by the sensors fluctuate around the true values by small random amounts. We assume that when there is no contaminant event, the concentration measurements collected by the M sensors at each time stamp are a series of i.i.d. multivariate normal random variables,

$$\boldsymbol{c}_t \overset{\text{i.i.d.}}{\sim} \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad t = 1, 2, \cdots,$$

where $\mathbf{0}$ represent a M-dimensional 0 vector and $\boldsymbol{\Sigma}$ is the spatial covariance matrix of the measurements. We often have enough reference data in the in-control case (no contaminant event) to estimate the mean and temporal correlation and to whiten the data, hence assuming $\boldsymbol{c_t}$ has zero mean and no temporal correlation when no contaminant event is reasonable. We also assume that $\boldsymbol{\Sigma}$ is known or has been estimated. When a contaminant event occurs, the mean of $\boldsymbol{c_t}$ is equal to the true concentration level of the contaminant in the water. In addition, temporal correlation exists due to the contaminant transport along the river.

In this case, a sensor network needs to detect a contaminant event with not only a low detection delay and a high reliability but also a low false alarm rate. A false alarm is defined as the case when there is no contaminant event but the sensors raise an alarm. We use the in-control average-run-length as the metric to quantify the frequency of a false alarm, which is the average number of time steps until an alarm is raised where there is no contaminant. For a specific $\boldsymbol{x}$, we use $\mathrm{ARL}_0(\boldsymbol{x})$ to denote the in-control average-run-length of the sensor network at $\boldsymbol{x}$. Then, the optimal sensor placement problem in the presence of measurement error is formulated as follows,

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathfrak{X}} \quad & \mathrm{E}[T(\boldsymbol{x})\big|R(\boldsymbol{x})=1] \\ \text{s.t.} \quad & \mathrm{E}[R(\boldsymbol{x})] \geq q, \quad q\in[0,1], \\ & \mathrm{ARL}_0(\boldsymbol{x}) \geq \mathrm{ARL}_{\text{target}}. \end{aligned} \tag{56}$$

The detection delay $T(\boldsymbol{x})$ not only depends on the sensor placement $\boldsymbol{x}$ but the detection statistic W_t as well. In this chapter, we use the Score Statistic for Spatio-Temporal surveillance ($\mathsf{S}^3\mathsf{T}$) proposed in Chapter 3 as the detection statistic. $\mathsf{S}^3\mathsf{T}$ statistic captures both spatial and temporal correlation information and is good at detecting weak signals. At each time stamp t, a sliding window of length ω is constructed to form the statistic. Define the following notations,

$$\boldsymbol{c}_{(t-\omega+1:t)} = [\boldsymbol{c}_{t-\omega+1}^{\mathsf{T}}, \cdots, \boldsymbol{c}_t^{\mathsf{T}}],$$

$$\boldsymbol{\Sigma}_\omega = \boldsymbol{I}_\omega \otimes \boldsymbol{\Sigma},$$

$$\boldsymbol{V}_\omega(\theta) = \boldsymbol{R}_\omega(\theta) \otimes \boldsymbol{\Sigma},$$

$$\boldsymbol{A}_\omega(\theta) = \boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{V}_\omega(\theta),$$

$$\boldsymbol{B}_\omega(\theta) = \boldsymbol{\Sigma}_\omega^{-1/2}\boldsymbol{V}_\omega(\theta)\boldsymbol{\Sigma}_\omega^{-1/2},$$

where $\boldsymbol{I}_\omega$ is a ω-by-ω identity matrix, $\otimes$ denotes the Kronecker produc and the matrix $\boldsymbol{R}_\omega(\theta) \in \mathbb{R}^{\omega\times\omega}$ captures the temporal correlation of the concentration measurements collected by the sensors when a contaminant event occurs. As suggested in [8] and [9], we assume that the temporal correlation follows a first-order vector autoregressive VAR(1)

model and hence $[\boldsymbol{R}_\omega(\theta)]_{i,j} = \theta^{|i-j|}, \forall i,j \in \{1,\cdots,\omega\}$. Then, the $\mathsf{S}^3\mathsf{T}$ statistic with window length ω is calculated as follows,

$$W_t(\omega) = \max_{\theta\in\Theta} \frac{\boldsymbol{c}_{(t-\omega+1:t)}^\mathsf{T} \boldsymbol{\Sigma}_\omega^{-1} \boldsymbol{V}_\omega(\theta) \boldsymbol{\Sigma}_\omega^{-1} \boldsymbol{c}_{(t-\omega+1:t)} - \mathrm{tr}\big(\boldsymbol{A}_\omega(\theta)\big)}{\sqrt{2\mathrm{tr}\big(\boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{V}_\omega(\theta)\boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{V}_\omega(\theta)\big)}}, \tag{57}$$

where $\mathrm{tr}\big(\,\cdot\,\big)$ denotes the trace of a matrix and Θ is a pre-specified parameter set.

To achieve a desired ARL_0 target, the threshold b needs to be adjusted. When spatial and temporal correlations exist in the data and $\mathsf{S}^3\mathsf{T}$ is used as the detection statistic, b may take different values for different $\boldsymbol{x}$ as the spatial and temporal covariances depend on the locations of the sensors. In this case, adjusting b is needed in every iteration of the optimization process. In practice, b can be adjusted based on simulations if the distribution of in-control data (i.e., when no contaminant event) is known or can be estimated. In the following, we briefly introduce an efficient simulation approach to adjusting b which is based on the fact that the distribution of run-length (denoted as RL, which is the number of time steps until a false alarm is raised) approximately follows exponential distribution. We first simulate m sequences of in-control data with a fixed length n, $\{\boldsymbol{c}_{i1},\cdots,\boldsymbol{c}_{in}\}_{i=1}^m$, where $\boldsymbol{c}_{ij} \in \mathbb{R}^M$, and construct m sequences of detection statistics based on the simulated data, $\{W_{i1},\cdots,W_{in}\}_{i=1}^m$. Next we obtain the maximum value of the detection statistics in each sequence, $\{W_i^\star\}_{i=1}^m$, where $W_i^\star = \max(W_{i1},\cdots,W_{in})$. For a specific b, we calculate $\hat{p} = \frac{\sum_{i=1}^m \mathbb{1}\{W_i^\star > b\}}{m}$, where $\mathbb{1}\{\cdot\}$ denotes an indicator function. Then, we have the following approximation,

$$\hat{p} \approx \mathbb{P}(\mathrm{RL} > n) \approx e^{-\lambda n},$$

where $\lambda = \mathrm{E}[\mathrm{RL}] = \mathrm{ARL}_0$. We can now obtain an approximated ARL_0 if a specific b value is used as the threshold. For a target ARL_0, we can find the appropriate b using a binary search. Such an approach is efficient because n needs not to be very large. We can obtain an accurate b using $n = \frac{\mathrm{ARL}_{\mathrm{target}}}{10}$.

4.3.3 Process simulation

In order to evaluate $\mathrm{E}[T(\boldsymbol{x})\big|R(\boldsymbol{x}) = 1]$ and $\mathrm{E}[R(\boldsymbol{x})]$ for different sensor placements on a river network, a simulation model that is capable of simulating hydrodynamics and contaminant transport is needed. The Storm Water Management Model (SWMM, [49]) developed by the United States Environmental Protection Agency is widely used in environmental engineering for water-related study. SWMM requires geologic, geometric and fundamental hydrodynamics data to construct a river network.

The random contaminant events are simulated by the SWMM model. The randomness of a contaminant event is due to the randomness of the location, intensity, duration and starting time of the spill as well as random rain events. For a specific sensor placement $\boldsymbol{x}$, a SWMM run is implemented as follows:

Step 1: Randomly generate the location, intensity, duration and starting time of a contaminant spill and rain pattern.

Step 2: Simulate contaminant transport along the river in consideration with the input data generated from Step 1.

Step 3: Evaluate $T(\boldsymbol{x})$ and $R(\boldsymbol{x})$ based on the concentration measurements collected at $\boldsymbol{x}$.

4.3.4 Wasserstein similarity metric

Bayesian optimization algorithms take advantage of the "smoothness" of a function, i.e., the function has similar values for similar decision variables. The similarities among different points in the search space are captured by the kernel function of the GP. An example of the commonly used kernel functions is the square-exponential kernel,

$$K(\boldsymbol{x}, \boldsymbol{x}^{'}) = \phi^2 \exp\Big\{ -\frac{1}{2} d\Big(\boldsymbol{x}, \boldsymbol{x}^{'}; \ell\Big)^2 \Big\},$$

where $\phi^2 > 0$ is the variance parameter and $d(\cdot, \cdot; \ell) \geq 0$ is a similarity metric parametrized by the length scale parameter $\ell > 0$. Smaller $d(\boldsymbol{x}, \boldsymbol{x}^{'}; \ell)$ indicates higher similarity between $\boldsymbol{x}$ and $\boldsymbol{x}^{'}$ in performance. If the decision variables are continuous variables or ordinal discrete

variables, a natural choice for d is the Euclidean distance,

$$d\Big(\boldsymbol{x}, \boldsymbol{x}^{'}; \ell\Big)^2 = \sum_{i=1}^{D} \frac{(\boldsymbol{x}_i - \boldsymbol{x}_i^{'})^2}{\ell}. \tag{58}$$

However, the decision variable $\boldsymbol{x}$ is a D-dimensional binary vector in problems (55) and (56), which makes Euclidean distance an inappropriate choice for the similarity metric. The reason is demonstrated in the following example. Figure 17 shows three different sensor placements (number of sensors $M = 2$) on a hypothetical river with $D = 6$ nodes. Denote the three sensor placements in Figure 17 (a), (b) and (c) as $\boldsymbol{x}_a$, $\boldsymbol{x}_b$ and $\boldsymbol{x}_c$, respectively. We have $\boldsymbol{x}_a = [1, 0, 0, 1, 0, 0]$, $\boldsymbol{x}_b = [0, 1, 1, 0, 0, 0]$ and $\boldsymbol{x}_c = [0, 0, 0, 0, 1, 1]$. The Euclidean distance between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ is 2 which is equal to the Euclidean distance between $\boldsymbol{x}_a$ and $\boldsymbol{x}_c$. However, as we can observe from Figure 17, from $\boldsymbol{x}_a$ to $\boldsymbol{x}_b$, we only need to move the two sensor from Node 1 to 2 and from Node 4 to 3 by small amounts, which will not greatly affect the performance. Hence, $\boldsymbol{x}_b$ is a much more similar placement to $\boldsymbol{x}_a$ than $\boldsymbol{x}_c$. Such difference is not captured by the Euclidean distance metric.

We propose to use the Wasserstein metric as the similarity metric. It is originally proposed as a measure of discrepancy between two distributions and is widely applied in image processing [50]. We adjust the metric to the setting of sensor placement on a river network. For a specific sensor placement, we first convert the D-dimensional binary vector $\boldsymbol{x}$ into a series of weights $\{w_i > 0\}_{i=1}^{D}$ by assigning $w_i = \frac{1}{M}$ if a sensor is placed at the ith node and $w_i = 0$ otherwise. Note that $\sum_{i=1}^{D} w_i = 1$. For two different sensor placements $\boldsymbol{x}$ and $\boldsymbol{x}^{'}$, and the corresponding weights $\{w_i > 0\}_{i=1}^{D}$ and $\{v_i > 0\}_{i=1}^{D}$, the Wasserstein distance, denoted as $d_w(\boldsymbol{x}, \boldsymbol{x}^{'})$, is intuitively the minimum amount of work required to move all sensors from placement $\boldsymbol{x}$ to $\boldsymbol{x}^{'}$. Formally, $d_w(\boldsymbol{x}, \boldsymbol{x}^{'})$ is equal to the optimum of the following linear program [50]:

$$\begin{aligned}
\min \quad & \textstyle\sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij} C_{ij} & \\
\text{s.t.} \quad & g_{ij} \geq 0, & 1 \leq i \leq M, 1 \leq j \leq M, \\
& \textstyle\sum_{j=1}^{M} g_{ij} = w_i, & 1 \leq i \leq M, \\
& \textstyle\sum_{i=1}^{M} g_{ij} = v_j, & 1 \leq i \leq M, \\
& \textstyle\sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij} = 1, &
\end{aligned} \tag{59}$$

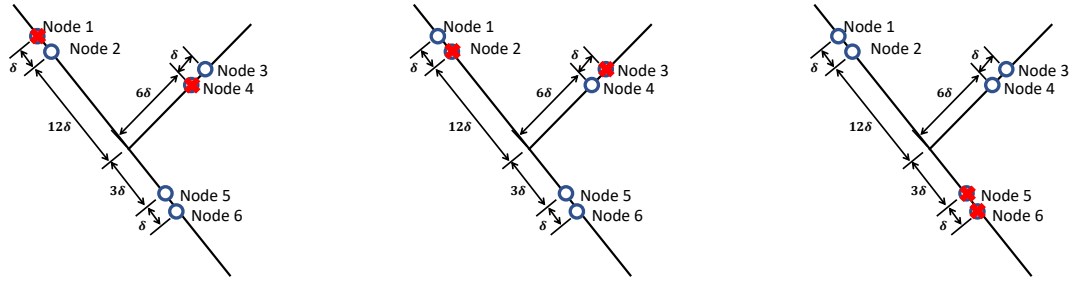


Figure 17: Three different sensor placements (number of sensors $M = 2$) on a hypothetical river with $D = 6$ nodes. Sensors are marked by the red crosses. The stream distances between each node are also marked on the plots.

where g_{ij} can be intuitively understood as the number of sensor moved from node i to j and C_{ij} is the cost of moving one sensor from node i to j. We use the stream distance, namely the shortest distance along the river between node i and j as C_{ij}. For the previous example, $d_w(\boldsymbol{x}_a, \boldsymbol{x}_b) = 2\delta$ and it is much smaller than $d_w(\boldsymbol{x}_a, \boldsymbol{x}_c) = 26\delta$.

4.4 Experiments

In this section, we solve problems (55) and (56) for the Altamaha River using the proposed methods. Backgrounds on the Altamaha River and SWMM model can be found in Section 2.5.1.

4.4.1 Simulation setup

In each SWMM run, 100 contaminant events are generated and each event is a single instantaneous spill at a different node. Note that the spills generated by the same SWMM run share the same rain pattern, but have different intensity and starting time. For a specific sensor placement $\boldsymbol{x}$, the output of each SWMM run are estimates of the conditional expected detection delay and reliability based on the 100 contaminant events.

4.4.2 Error-free sensors

We first assume that the sensors are error-free and consider problem (55).

4.4.2.1 Settings

In this case, false alarms are not of our concern and the threshold b is chosen arbitrarily. We adopt the same setting in [41]: $b \in \{0.05, 0.0001\}$ (milligram/litter). Input parameters for CSCBO include h_1 and h_2 which control the way CS_1 and CS_2 are constructed, as well as k in the stopping criteria which is the minimum number of consecutive iterations where the same point is reported as the optimal feasible solution. We set $h_1 = 0.025$, $h_2 = 0.5$ and $k = 50$. We use independent GPs in the algorithm, as the benefit of MTGP is not obvious due to the high-dimensional nature of the problem. A square-exponential kernel as defined in (58) is used as the kernel function of the GP for $\mathrm{E}[R(\boldsymbol{x})]$. The kernel function of the GP for $\mathrm{E}[T(\boldsymbol{x})|R(\boldsymbol{x}) = 1]$ is a Matérn 5/2 model, as defined in the following,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \phi^2\Big(1 + \frac{\sqrt{5}d_w(\boldsymbol{x}, \boldsymbol{x}')}{\ell} + \frac{5d_w(\boldsymbol{x}, \boldsymbol{x}')^2}{3\ell^2}\Big)\exp\Big\{-\frac{\sqrt{5}d_w(\boldsymbol{x}, \boldsymbol{x}')}{\ell}\Big\},$$

where $\ell > 0$ is the length scale parameter. Note that we do not need to specify the prior mean, prior variance and length scale parameter ℓ for the GPs as the maximum likelihood estimates of these parameters [45] are used. The minimum reliability level is set as $q = 0.9$. We test two possible values for the number of sensors, $M \in \{5, 7\}$.

Implementing a Bayesian optimization algorithm is challenging when the size of the search space $\mathfrak{X}$ is large. In our problem, the size of $\mathfrak{X}$ grows exponentially in the number of possible sensor location on the river network D when the number of sensors M is fixed. For example, if $M = 5$ and the number of possible sensor location $D = 100$, we have $|\mathfrak{X}| = \binom{100}{5} = 75287520$, which makes implementing CSCBO on a personal desktop impossible. To speed up the computation, we exclude the nodes on the most upstream points in the search as placing sensors on these nodes is obviously not beneficial. We limit the number of possible sensor locations to $D = 50$ for $M = 5$ and $D = 30$ for $M = 7$.

In the experiment, we compare CSCBO with NP + PFM [41]. Three performance measures are used in the comparison: (i) estimated conditional expected detection delay (ECEDD) and (ii) estimated reliability (ER) of the optimal feasible solution found by the algorithm as well as (iii) the number of SWMM runs required until termination of the algorithm (NUM).

Table 10: Performance metrics of the optimal feasible solutions found by CSCBO and NP + PFM. The unit of ECEDD is hour.

		CSCBO			NP + PFM		
M	b	ECEDD	ER	NUM	ECEDD	ER	NUM
5	0.05	59.16	0.901	205	55.54	0.902	614
5	0.0001	49.01	0.930	95	46.13	0.930	154
7	0.05	45.49	0.903	237	44.47	0.916	800
7	0.0001	38.48	0.930	113	37.05	0.930	205

4.4.2.2 Results

Figure 18 shows the optimal feasible solution found by CSCBO in circles and NP + PFM in triangles for different settings. Table 10 presents the corresponding performance metrics of the two methods. From the results, we observe that all the solutions found by CSCBO are feasible, indicating confidence-set is a effective method to deal with optimization problems with black-box function constraints. In addition, CSCBO is able to find competitive sub-optimal solutions using much fewer number of SWMM runs than NP + PFM. Considering that NP + PFM visits multiple solutions in each iteration (200 solutions per iteration in the experiment), CSCBO converges much more quickly than NP + PFM in terms of the number of solutions visited. Such advantage of CSCBO leads to tremendous savings on computation resources considering SWMM is a highly computational-intensive model and each SWMM run takes around 2 hours on a personal desktop with Intel Core I5 CPU.

It turns out that CSCBO tends to converge to a local optimum in problem (55). We find that the algorithm has difficulty finding solutions that are better than the ones reported in Figure 18 when the algorithm is allowed to run more iterations. The difficulty mainly arises from the high dimensional nature of the problem. BO algorithms usually perform well on problems with low dimensionality and continuous decision variables but suffer in high-dimensional problems with nominal variables. Although the Wasserstein similarity metric discussed in Section 4.3.4 enables CSCBO to deal with a problem with D-dimensional binary variables, it is still difficult for the algorithm to identify the true optimal feasible solution given the huge search space. Hence, CSCBO is a favorable method when computation

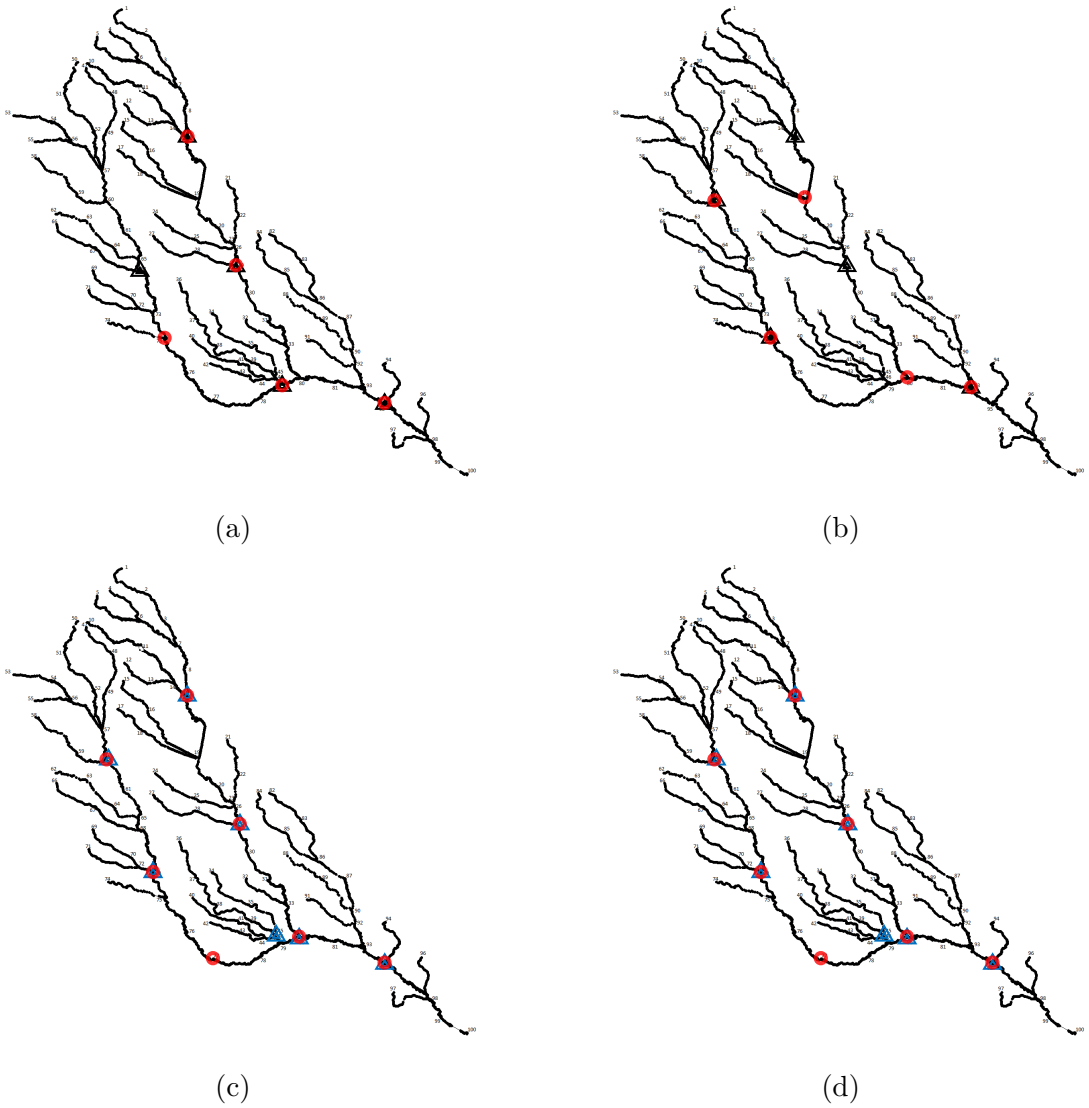


Figure 18: Optimal feasible solutions found by CSCBO (circle) and NP + PFM (triangle): (a) $M = 5$ and $b = 0.05$, (b) $M = 5$ and $b = 0.0001$, (c) $M = 7$ and $b = 0.05$ and (d) $M = 7$ and $b = 0.0001$.

budget is tight and a local optimal feasible solution with little sacrifice on the objective function is acceptable.

4.4.3 Sensors with measurement error

Here we assume that sensor measurement error exists and apply CSCBO and $\mathsf{S}^3\mathsf{T}$ statistic to solving problem (56).

Table 11: ECEDD (in hours) of sensor placements marked by circles in Figure 18(a) and (c) using different detection statistics. Numbers in parentheses are standard errors.

M	ζ_1	$\mathsf{S}^3\mathsf{T}$	Shewhart	max-CUSUM
5	0.005	71.13 (0.21)	81.70 (0.34)	73.21 (0.20)
5	0.01	72.91 (0.24)	85.18 (0.39)	75.65 (0.23)
7	0.005	56.48 (0.20)	66.77 (0.31)	58.43 (0.16)
7	0.01	58.14 (0.22)	75.59 (0.41)	61.57 (0.22)

4.4.3.1 Settings

We adopt the "tail-up" model proposed in [69] (see 2.5.2 for details) with an exponential correlation function for the spatial correlation of the concentration measurements collected by the sensors. The concentration measurements follow the same spatial correlation structure before and after a contaminant event. The maximum likelihood estimate $\hat{\zeta}_2 = 0.68$ is obtained based on the data simulated by the SWMM model. We set the marginal variance of the measurement error $\zeta_1 \in \{0.005, 0.01\}$. $\text{ARL}_{\text{target}}$ is set to 10,000 which is approximately equivalent to 100 days as the inter-reporting time is set to 15 minutes in the SWMM model. The $\mathsf{S}^3\mathsf{T}$ statistic defined in (57) with a window length $\omega = 20$ is used as the detection statistic. The search space of the parameter θ in a VAR(1) model is set to $\{0.1, 0.2, \cdots, 0.9\}$. For CSCBO, we use the same setting as in Section 4.4.3.1. The minimum requirement for reliability is set as $q = 0.9$.

4.4.3.2 Comparison on detection delay

To demonstrate the advantage of the $\mathsf{S}^3\mathsf{T}$ statistic, we compare $\mathsf{S}^3\mathsf{T}$ with two other detection statistics in terms of conditional expected detection delay for fixed sets of sensors locations when the target ARL_0 is set to a pre-specified level: (i) a simple detection statistic defined in (52) which we refer to as the Shewhart statistic and (ii) a max-CUSUM statistic defined as follows,

$$\text{Cusum}_{t,i} = \max(0,\ \text{Cusum}_{t-1,i} + c_i - \kappa\sqrt{\zeta_1}), \quad i = 1, \cdots, M, \tag{60}$$

and

$$W_t = \max_{i \in \{1, \cdots, M\}} \text{Cusum}_{t,i}, \tag{61}$$

Table 12: Simulated ARL_0 using $\mathsf{S}^3\mathsf{T}$ when no contaminant event. $\text{ARL}_{\text{target}} = 10000$.

M	ζ_1	b	ARL_0
5	0.005	4.989	10276
5	0.01	5.025	10426
7	0.005	4.788	9904
7	0.01	4.847	10143

where $\kappa > 0$ is a constant and we set $\kappa = 0.1$ in the experiment. Here we assume that the marginal variance of the sensor measurement error ζ_1 is known in the max-CUSUM statistic, while in practice, it should be estimated from data. The threshold b for the three statistics are adjusted via simulations so that the ARLs are approximately 10,000.

We use the two sensor placements marked by circles in Figure 18(a) ($M = 5$) and (c) ($M = 7$) in the experiments. Note that these solutions are not necessarily feasible for problem (56) and the focus of this experiment is on the performance of the $\mathsf{S}^3\mathsf{T}$ statistic alone. We compare the ECEDDs of each sensor placement using different detection statistics. The ECEDDs and the corresponding standard error are obtained based on 1000 SWMM runs and presented in Table 11. As we can observe, $\mathsf{S}^3\mathsf{T}$ outperforms the other two statistics in all experiments and achieves 7 - 12 hours shorter conditional expected detection delay comparing to the Shewhart statistic and 1 - 3 hours comparing to the max-CUSUM statistic. The advantage is more obvious when the variance of measurement error is high ($\zeta_1 = 0.01$). Such an advantage of $\mathsf{S}^3\mathsf{T}$ is due to its capacity of capturing both spatial and temporal correlation in the data.

4.4.3.3 Adjusting threshold b

Here we validate the accuracy of the simulation approach for adjusting b discussed in Section 4.3.2. We test the two sensor placements marked by circles in Figure 18(a) ($M = 5$) and (c) ($M = 7$). In the experiment, we first adjust b using the discussed approach with $\text{ARL}_{\text{target}} = 10000$ and then simulate in-control data to obtain the actual ARL_0. The number of replications for estimating ARL_0 is 2000. The results is presented in Table 12 where we can observe that the adjusted b values are accurate enough for the detection

Table 13: Performance metrics of the optimal feasible solution found by CSCBO with $\mathsf{S}^3\mathsf{T}$ as the detection statistic. The unit of ECEDD is hour.

M	ζ_1	ECEDD	ER	false_alarm_rate	NUM
5	0.01	70.60	0.931	0.2%	147
5	0.005	64.36	0.924	0.26%	181
7	0.01	62.46	0.934	0.32%	123
7	0.005	57.93	0.925	0.31%	253

statistic to achieve the target ARL_0.

4.4.3.4 Optimal solution in the presence of measurement error

Now we combine CSCBO and $\mathsf{S}^3\mathsf{T}$ to solve problem (56). In each iteration of the optimization, the threshold b is adjusted so that the ARL_0 is approximately 10,000. Figure 19 presents the optimal feasible solutions found by CSCBO with $\mathsf{S}^3\mathsf{T}$ as the detection statistics for $M = 5$ and $M = 7$. Table 13 reports the performance metrics of these solutions based on 1000 SWMM runs. Note that in the experiments, a false alarm is defined as an alarm raised by the sensors before the actual starting time of a simulated contaminant event. We calculate false alarm rates as the number of false alarms divided by the total number of simulated contaminant events. From Table 13, we first observe that all solutions have estimated reliability higher than $q = 0.9$, and hence the reported solutions are feasible. The false alarm rates are around 0.3% in all cases, indicating that by controlling the ARL_0 of sensor networks, the occurrence of false alarms can be controlled at a very low level. In addition, we can observe from Figure 19 that the optimal sensor placements found by the algorithm do not differ much for different levels of the variance of measurement error. This means that the optimal sensor network design identified by the combined procedure of CSCBO and $\mathsf{S}^3\mathsf{T}$ is robust to different level of variances of the sensor measurement error.

4.5 Conclusion

In this chapter, we study the problem of optimal sensor network design in the presence of random measurement error for water quality monitoring on river systems. The spatial and temporal correlation in the sensor measurement is take into consideration. We formulate

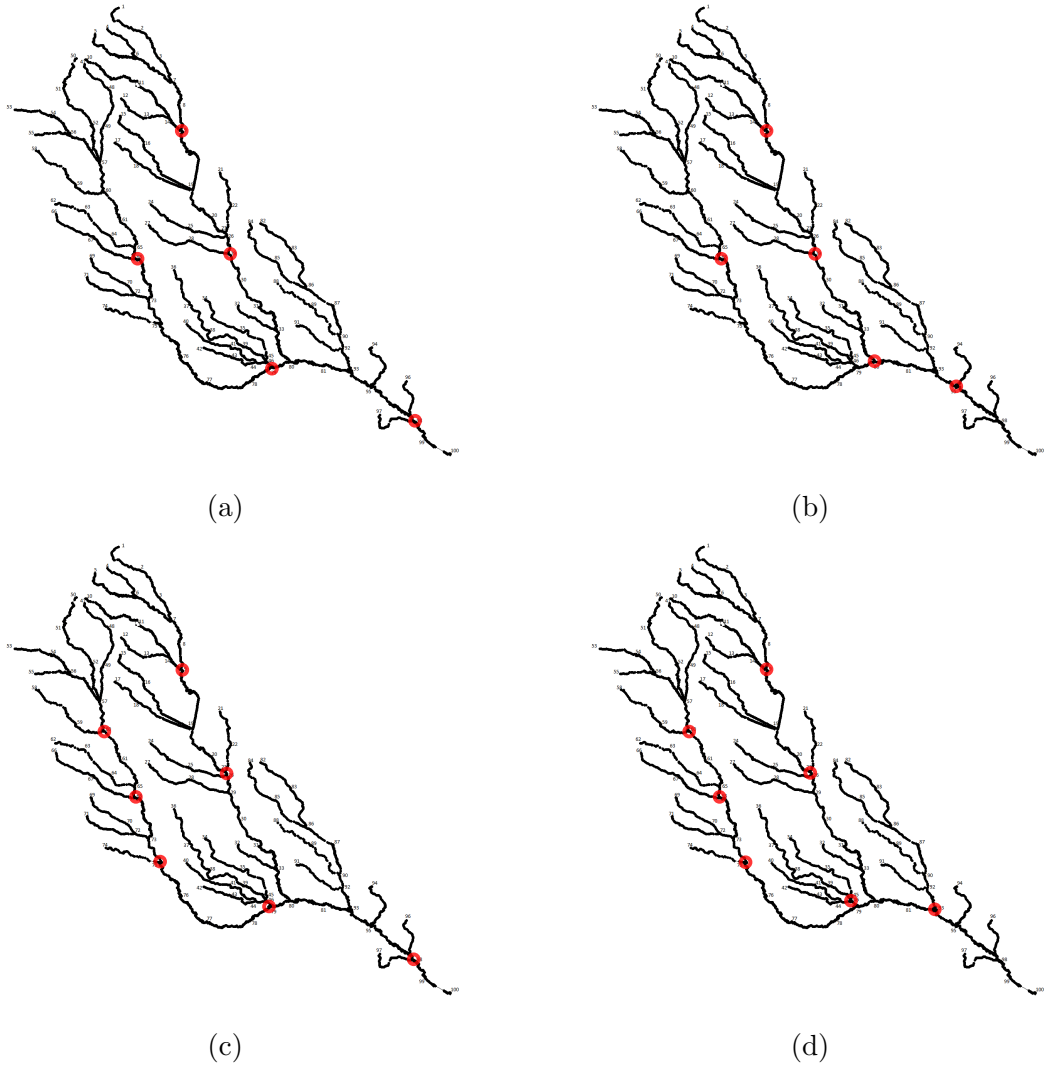


Figure 19: Optimal feasible solutions found by CSCBO with $\mathsf{S}^3\mathsf{T}$ as the detection statistic: (a) $M = 5$ and $\zeta_1 = 0.01$, (b) $M = 5$ and $\zeta_1 = 0.005$, (c) $M = 7$ and $\zeta_1 = 0.01$ and (d) $M = 7$ and $\zeta_1 = 0.005$.

the problem as a joint problem of constrained black-box function optimization and spatio-temporal change-point detection. The proposed algorithm CSCBO with a Wasserstein similarity metric is demonstrated to converge quickly and is able to find competitive feasible suboptimal sensor placements. By combining CSCBO with the $\mathsf{S}^3\mathsf{T}$ statistic, we identify sensor network designs which are shown to have low conditional detection delay, required reliability and low false-alarm rates. In addition, the identified sensor networks are robust to different levels of measurement error variances.

CHAPTER V

DISCUSSION AND CONCLUSIONS

In Chapter 2, we discuss dimension reduction via spatial scanning. The key question we aim to answer is how much do we lose in terms of detection performance by using reduced-dimension spatial scanning comparing to methods using full observation vectors. Our studies show that the performance loss of the RD approach is within the acceptable range. Considering that the RD spatial scanning enjoys all the computation and implementation benefits, it should be a preferable method for practitioners.

In Chapter 3, we propose the $\mathsf{S}^3\mathsf{T}$ statistic to detect the emergence of a spatially and temporally correlated signal from noisy background. The proposed statistic jointly captures the spatial and temporal correlations of the signal. Numerical studies based on simulated and real data demonstrate that $\mathsf{S}^3\mathsf{T}$ outperforms the baseline methods. Our results also show that the advantage of $\mathsf{S}^3\mathsf{T}$ is more obvious when the signal to be detected is weak.

In Chapter 4, we study the problem of optimal sensor network design, which is formulated as a joint problem of constrained black-box function optimization and spatio-temporal change-point detection. We propose the Confidence-Set based Constrained Bayesian Optimization (CSCBO) algorithm for general constrained black-box function optimization problems and extend the algorithm to tackle with high-dimensional binary decision variables by using the Wasserstein similarity metric. Experiments reveal that confidence-set is an effective and flexible method to handle black-box function constraints. In addition, CSCBO converges much quicker than NP+PFM in the water monitoring network application and can identify competitive local optimal solutions with little loss on the conditional expected detection delay. Finally, sensor network designs identified by combining CSCBO and $\mathsf{S}^3\mathsf{T}$ are shown to be robust to sensor measurement errors.

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Derivation of (12)

First note that $d_0 < 0$ in all cases and we assume that $d_{c,r} > 0$. Given target ARL_0, d_0, $d_{c,r}$, Ω_0^2 and $\Omega_{c,r}^2$, we need to find H and ARL_1 that satisfy the following approximations from (11):

$$\text{ARL}_0 \approx \frac{\Omega_0^2}{2d_0^2}\left\{\exp\left[-\frac{2d_0(H+1.166\Omega_0)}{\Omega_0^2}\right] - 1 + \frac{2d_0(H+1.166\Omega_0)}{\Omega_0^2}\right\};$$

and

$$\text{ARL}_1 \approx \frac{\Omega_{c,r}^2}{2d_{c,r}^2}\left\{\exp\left[-\frac{2d_{c,r}(H+1.166\Omega_{c,r})}{\Omega_{c,r}^2}\right] - 1 + \frac{2d_{c,r}(H+1.166\Omega_{c,r})}{\Omega_{c,r}^2}\right\}.$$

From the above two approximations, we get

$$\frac{2d_0^2}{\Omega_0^2}\text{ARL}_0 + 1 \approx \left\{\exp\left[-\frac{2d_0(H+1.166\Omega_0)}{\Omega_0^2}\right] + \frac{2d_0(H+1.166\Omega_0)}{\Omega_0^2}\right\}; \tag{62}$$

and

$$\frac{2d_{c,r}^2}{\Omega_{c,r}^2}\text{ARL}_1 + 1 \approx \left\{\exp\left[-\frac{2d_{c,r}(H+1.166\Omega_{c,r})}{\Omega_{c,r}^2}\right] + \frac{2d_{c,r}(H+1.166\Omega_{c,r})}{\Omega_{c,r}^2}\right\}. \tag{63}$$

Let $\eta_0 = \frac{2d_0^2}{\Omega_0^2}\text{ARL}_0 + 1$ and $\eta_1 = \frac{2d_{c,r}^2}{\Omega_{c,r}^2}\text{ARL}_1 + 1$. In general, a solution x to an equation $e^x + x = c$ for a constant c can be expressed using the Lambert W function [10]. Then using the Lambert W function, we get

$$\frac{2d_0(H+1.166\Omega_0)}{\Omega_0^2} = W_{-1}(-e^{-\eta_0}) + \eta_0 \quad \text{from (62)}, \tag{64}$$

$$\frac{2d_{c,r}(H+1.166\Omega_{c,r})}{\Omega_{c,r}^2} = W_0(-e^{-\eta_1}) + \eta_1 \quad \text{from (63)}, \tag{65}$$

where $W_{-1}(\cdot)$ and $W_0(\cdot)$ denote the two branches of the Lambert W function [10]. Note that we need $W_{-1}(\cdot)$ for (64) because the LHS is a negative number due to $d_0 < 0$ while we need $W_0(\cdot)$ for (65) because the LHS is a positive number due to $d_{c,r} > 0$.

Solving (64) and (65) for H results in

$$H = \frac{\Omega_0^2}{2d_0}(W_{-1}(-e^{-\eta_0}) + \eta_0) - 1.166\Omega_0$$

and

$$H = \frac{\Omega_{c,r}^2}{2d_{c,r}}(W_0(-e^{-\eta_1}) + \eta_1) - 1.166\Omega_{c,r}.$$

Then we obtain

$$\frac{\Omega_0^2}{2d_0}(W_{-1}(-e^{-\eta_0}) + \eta_0) - 1.166\Omega_0 = \frac{\Omega_{c,r}^2}{2d_{c,r}}(W_0(-e^{-\eta_1}) + \eta_1) - 1.166\Omega_{c,r}.$$

Let $\epsilon_{\eta_0} = -W_{-1}(-e^{-\eta_0}) - \eta_0$, which is a positive constant number. Then

$$\frac{\Omega_{c,r}^2}{2d_{c,r}}(W_0(-e^{-\eta_1}) + \eta_1) = -\frac{\Omega_0^2}{2d_0}\epsilon_{\eta_0} - 1.166(\Omega_0 - \Omega_{c,r}). \tag{66}$$

The LHS of (66) is

$$\frac{\Omega_{c,r}^2}{2d_{c,r}}\left(W_0(-e^{-\eta_1}) + \frac{2d_{c,r}^2}{\Omega_{c,r}^2}\mathrm{ARL}_1 + 1\right).$$

Note that $-1 \leq W_0(-e^{-\eta_1}) \leq 0$ and $1 + W_0(-e^{-\eta_1}) \approx 0$. Then we finally obtain (12) as follows,

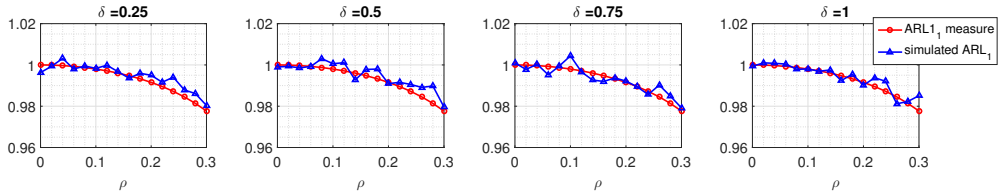
$$\mathrm{ARL}_1 \approx -\frac{\Omega_0^2}{2d_0 d_{c,r}}\epsilon_{\eta_0} - 1.166\frac{1}{d_{c,r}}(\Omega_0 - \Omega_{c,r}).$$

A.2 Figures and Results for Section 2.3.2.1

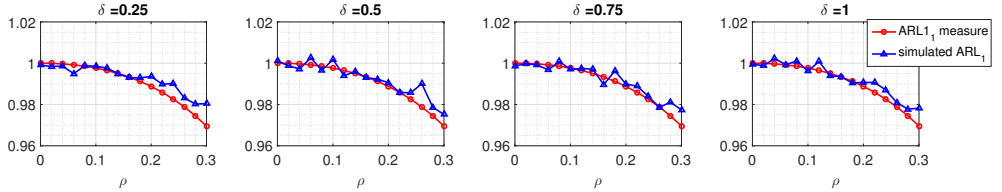
Numerical Examples with Known Center and Known Radius: Here we conduct simulation experiments assuming the shift cluster is known and show that simulation results match the ARL_1 measures very well.

Consider two different spatial covariance matrices: (i) tridiagonal matrix $\mathbf{\Sigma}_1(\rho)$ and (ii) $\mathbf{\Sigma}_2(\rho) \in R_{p\times p}$ with $[\mathbf{\Sigma}_2(\rho)]_{i,j} = \rho^{|i-j|},\ \forall i,j \in \{1,\cdots,p\}$, where the correlation decays polynomially. We use $p = 49$, $\tilde{p} = 5$ and $\rho \in \{0,\ 0.02,\ 0.04,\cdots,0.3\}$. For the out-of-control state, homogeneous shifts (shifts of all affected locations in the cluster have same magnitude) with magnitudes $\delta = 0.25,\ 0.5,\ 0.75$ and 1 are tested. The targeted ARL_0 is fixed to 1000 in all the cases.

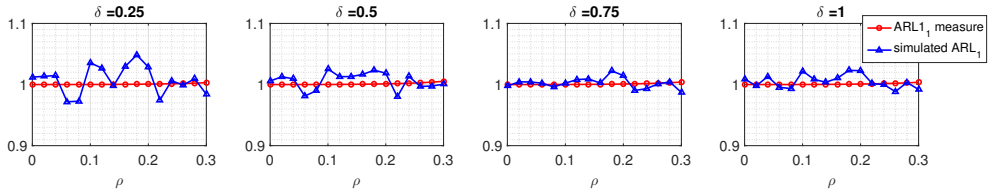
We report the ratio of the simulated ARL_1 of a F-MCUSUM chart to a RD-MCUSUM chart. The result is shown in Figure 20. In the same plot, we also present the ratio of the



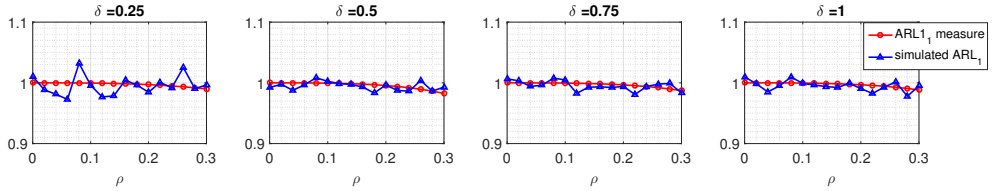
(a) LR based charts with Σ_1.



(b) LR based charts with Σ_2.



(c) T^2 based charts with Σ_1.



(d) T^2 based charts with Σ_2.

Figure 20: Ratio of simulated ARL_1's (blue) and ratio of ARL_1 measures (red) in the known shift cluster case.

ARL_1 measures, $m_{\text{LR}}/\tilde{m}_{\text{LR}}$ and $m_{T^2}/\tilde{m}_{T^2}$. Note that for T^2 based charts, when the shift magnitude $\delta = 0.25$, the out-of-control drift $d_{c,r}$ is negative. Hence, the ARL_1 measures are not applicable in this case. Instead, we numerically search for the control limit H given a ARL_0, obtain d_0 and Ω_0 using Formula (11), and then substitute H into (11) with $d_{c,r}$ and $\Omega_{c,r}$ to obtain an approximate ARL_1 for both T^2-F-MCUSUM and T^2-RD-MCUSUM charts.

From Figure 20, we see that simulation results match very well with the ARL_1 measures. This indicates that the ARL_1 measure is indeed a good measure for detection performance comparison. Moreover, we observe that the F-MCUSUM and RD-MCUSUM charts have similar detection performances in the known shift cluster case: (i) for LR based methods,

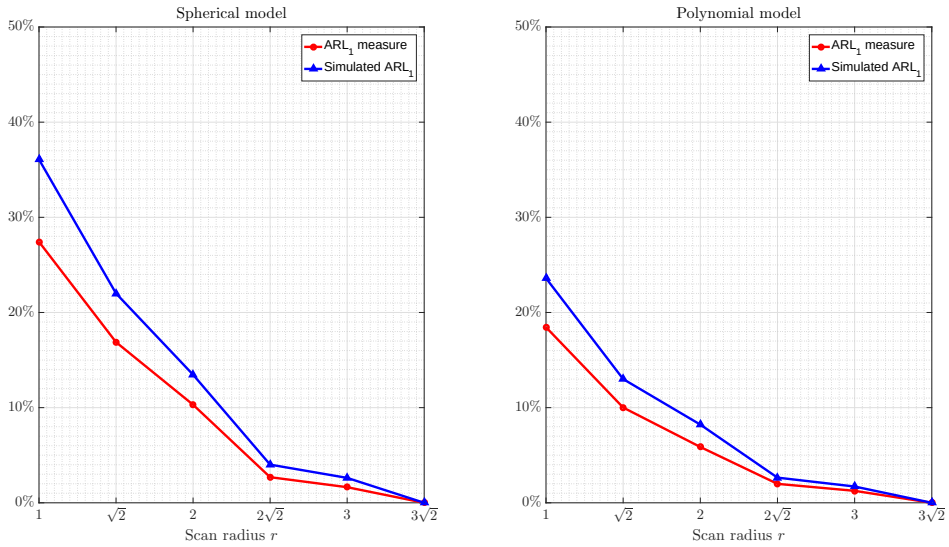


Figure 21: Performance loss based on ARL_1 measure and simulated ARL_1.

the F-MCUSUM chart slightly outperform RD-MCUSUM chart since all the ratios is less than but very close to 1; (ii) for T^2 based methods, depending on different spatial correlation structures, $m_{T^2}/\tilde{m}_{T^2}$ is either slightly greater or smaller than 1, however, we can hardly see the difference in performance from simulation results.

A.3 Decaying Shift

Here we consider "bell" shaped shift signals. Such a shift signal occurs at a center $c \in P$, affect all locations or sensors, and the shift magnitude decays with the distances from c. In this case, reduced-dimension charts are destined to lose performance, since each reduced-dimension observation can only capture a portion of the signal energy. The smaller the scanning radius, the more dimensionality reduction we achieve since each cluster has fewer data streams, but there is a great loss of performance.

We present a simple example in the presence of such a decaying shift. This example demonstrates one practical use of our theory: analytically choose the smallest scan radius r allowed based on the ARL_1 measure, given a maximum acceptable performance loss. Here we use percentage loss as the metric:

$$\frac{\mathrm{ARL}_1 \text{ of an RD-MCUSUM chart} - \mathrm{ARL}_1 \text{ of a F-MCUSUM chart}}{\mathrm{ARL}_1 \text{ of an F-MCUSUM chart}} \times 100\%,$$

and compare it with the ratio with approximated ARL_1 measures defined as $(\tilde{m}_{LR} - m_{LR})/m_{LR}$ using (13). Consider a shift signal with a magnitude δ at the center c, and the magnitude at other locations $d \in P$ is given by $\delta\theta^{||q_c - q_d||}$, where $\theta \in (0, 1)$ is the delay rate. Assume a monitoring area of dimension $p = 7 \times 7$. We run the control charts on the spherical model and the polynomial model with correlation parameter $\rho = 0.2$. We let $\delta = 1$, $\theta = 0.5$ and target $\text{ARL}_0 = 1000$. Our goal is to determine a smallest scan radius r (i.e., maximum dimensionality reduction), such that the performance loss in ARL_1 is no more than, say, 10% when using reduced dimension vectors. We may choose r from $\{0, 1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}\}$. Note that $r = 3\sqrt{2}$ corresponds to using full observation vectors, and $r = 0$ means each scanning region has only one sensor. In Figure 21, the blue curve shows the performance loss in terms of the simulated ARL_1, while the red curve shows $(\tilde{m}_{LR} - m_{LR})/m_{LR}$. From Figure 21, we conclude that we need $r = 2$ or $2\sqrt{2}$ under Σ_3 and $r = \sqrt{2}$ or 2 under Σ_4 to achieve no more than 10% loss in ARL_1.

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Derivation of $\frac{\partial \ell}{\partial \mu}\big|_{\mu=0,\gamma=0}$.

The following propositions are used in the derivation of $\frac{\partial \ell}{\partial \boldsymbol{\mu}}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0}$.

Proposition B.1.1. *Let $\boldsymbol{M}(t)$ be a nonsingular square matrix whose elements are functions of a scalar parameter α. Then,*

$$\frac{\partial \boldsymbol{M}(t)^{-1}}{\partial \alpha} = -\boldsymbol{M}(t)^{-1}\frac{\partial \boldsymbol{M}(t)}{\partial \alpha}\boldsymbol{M}(t)^{-1}.$$

Proposition B.1.2. *Let $\boldsymbol{M}(t)$ be a nonsingular square matrix whose elements are functions of a scalar parameter α. Then,*

$$\frac{\partial |\boldsymbol{M}(t)|}{\partial \alpha} = |\boldsymbol{M}(t)|\mathrm{tr}\bigg(\boldsymbol{M}(t)^{-1}\frac{\partial \boldsymbol{M}(t)}{\partial \alpha}\bigg).$$

By Proposition B.1.1, we can calculate,

$$\begin{aligned}\frac{\log\big|\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau\big|}{\partial\gamma}\bigg|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0} &= \frac{1}{\big|\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau\big|}\big|\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau\big|\mathrm{tr}\Big((\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau)^{-1}\boldsymbol{V}_\tau(\theta)\Big)\bigg|_{\gamma=0}\\ &= \mathrm{tr}\big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big).\end{aligned}$$

For convenience, here we use $\boldsymbol{y}$ and $\boldsymbol{\mu}$ to denote $\boldsymbol{y}_{(k+1:N)}$ and $\boldsymbol{\mu}_{(k+1:N)}$. By Proposition B.1.2, we have,

$$\begin{aligned}\frac{\partial(\boldsymbol{y}-\boldsymbol{\mu})^\intercal(\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau)^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\partial\gamma}\bigg|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0} &= (\boldsymbol{y}-\boldsymbol{\mu})^\intercal\frac{\partial(\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau)^{-1}}{\partial\gamma}(\boldsymbol{y}-\boldsymbol{\mu})\bigg|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0}\\ &= -\boldsymbol{y}^\intercal(\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau)^{-1}\boldsymbol{V}_\tau(\theta)(\gamma \boldsymbol{V}_\tau(\theta)+\boldsymbol{\Sigma}_\tau)^{-1}\boldsymbol{y}\bigg|_{\gamma=0}\\ &= -\boldsymbol{y}^\intercal\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y}.\end{aligned}$$

Hence we have,

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}}\big|_{\boldsymbol{\mu}=\mathbf{0},\gamma=0} = -\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big)+\frac{1}{2}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{y},$$

as appeared in equation (24).

B.2 Derivation of $\mathrm{Var}\big[S(\tau,\theta)\big]$*.*

Here we calculate the variance of the statistic $S(\tau,\theta)$ defined in (25). For convenience, we use $\boldsymbol{y}$ to denote $\boldsymbol{y}_{(k+1:N)}$, use $\boldsymbol{V}$ to denote the matrix $\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}$, use $\boldsymbol{\Sigma}$ to denote $\boldsymbol{\Sigma}_{\boldsymbol{\tau}}$ and use c and d to denote $c(\tau,\theta)$ and $d(\tau,\theta)$, respectively. Then we can write,

$$S(\tau,\theta) = \frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^2}{d} + \boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

We have that $\mathrm{E}[S] = p\tau + 1$, and

$$\mathrm{Var}\big[S(\tau,\theta)\big] = \mathrm{E}[S^2] - \mathrm{E}[S]^2.$$

In the following, we calculate $\mathrm{E}[S^2]$.

$$\begin{aligned}\mathrm{E}[S^2] &= \mathrm{E}\left[\left(\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^2}{d} + \boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)^2\right] \\ &= \mathrm{E}\left[\left(\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)^2\right] + 2\mathrm{E}\left[\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^2}{d}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] + \mathrm{E}\left[\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^4}{d^2}\right].\end{aligned} \tag{67}$$

The first term can be calculated as,

$$\mathrm{E}\left[\left(\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)^2\right] = p^2\tau^2 + 2p\tau. \tag{68}$$

We then expand the second term,

$$\begin{aligned}\mathrm{E}\left[\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^2}{d}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] &= \frac{1}{d}\mathrm{E}\left[\left(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y}\right)^2\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] \\ &\quad - \frac{2c}{d}\mathrm{E}\left[\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] + \frac{c^2}{d}\mathrm{E}\left[\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right].\end{aligned}$$

We calculate the three expectations separately:

$$\mathrm{E}\left[\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] = p\tau.$$

$$\mathrm{E}\left[\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] = (p\tau + 2)c.$$

$$\mathrm{E}\left[\left(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y}\right)^2\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] = (p\tau + 4)(2d + c^2).$$

Combining we get,

$$\mathrm{E}\left[\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y} - c)^2}{d}\boldsymbol{y}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] = 2p\tau + 4. \tag{69}$$

Next we calculate the last term in (67),

$$\mathrm{E}\left[\frac{(\boldsymbol{y}^\intercal \boldsymbol{V}\boldsymbol{y}-c)^4}{d^2}\right] = 3-2p\tau-24\frac{c}{d^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\Big)+\frac{48}{d^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\Big). \tag{70}$$

Note that the tedious calculation steps for (70) are omitted here.

Combining (68), (69) and (70), we can obtain,

$$\mathrm{Var}\big[S(\tau,\theta)\big] = 2p\tau+10-24\frac{c}{d^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\Big)+\frac{48}{d^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\Big).$$

B.3 Derivation of the cumulant generating function of W.

Here we present the derivation of the cumulant generating function of $W(\tau,\theta)$ under the null hypothesis, i.e. equation (34).

Let $\boldsymbol{z}=\boldsymbol{\Sigma}_\tau^{-\frac{1}{2}}\boldsymbol{y}_{(k+1:N)}$. Under the null hypothesis, we have $\boldsymbol{z}\sim\mathcal{N}\big(\boldsymbol{0},\boldsymbol{I}_{p\tau}\big)$. For convenience, here we use $\boldsymbol{B}$ to denote the $p\tau$ by $p\tau$ matrix $\boldsymbol{\Sigma}_\tau^{-\frac{1}{2}}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-\frac{1}{2}}$, and use c and d to denote $c(\tau,\theta)$ and $d(\tau,\theta)$, respectively. Then, we have

$$W(\tau,\theta)=\frac{\boldsymbol{z}^\intercal\boldsymbol{B}\boldsymbol{z}-c}{\sqrt{d}}.$$

Under the null hypothesis, the cumulant generating function of $W(\tau,\theta)$ can be calculated as

$$\begin{aligned}
\psi(\xi) &= \log\mathrm{E}[\exp(\xi W(\tau,\theta))] = \log\mathrm{E}\Big[\exp\Big(\xi\Big(\frac{\boldsymbol{z}^\intercal\boldsymbol{B}\boldsymbol{z}-c}{\sqrt{d}}\Big)\Big)\Big]\\
&= -\xi\frac{c}{\sqrt{d}}+\log\mathrm{E}\Big[\exp\Big(\frac{\xi\boldsymbol{z}^\intercal\boldsymbol{B}\boldsymbol{z}}{\sqrt{d}}\Big)\Big]\\
&= -\xi\frac{c}{\sqrt{d}}+\log\int_{\boldsymbol{z}}\exp\Big(\frac{\xi\boldsymbol{z}^\intercal\boldsymbol{B}\boldsymbol{z}}{\sqrt{d}}\Big)\frac{1}{(2\pi)^{\frac{p\tau}{2}}}\exp\Big(-\frac{1}{2}\boldsymbol{z}^\intercal\boldsymbol{z}\Big)d\boldsymbol{z}\\
&= -\xi\frac{c}{\sqrt{d}}+\log\int_{\boldsymbol{z}}\frac{1}{(2\pi)^{\frac{p\tau}{2}}}\exp\Big(-\frac{1}{2}\boldsymbol{z}^\intercal\Big(\boldsymbol{I}_{p\tau}-\frac{2\xi\boldsymbol{B}}{\sqrt{d}}\Big)\boldsymbol{z}\Big)d\boldsymbol{z}\\
&= -\xi\frac{c}{\sqrt{d}}+\log\Big|\boldsymbol{I}_{p\tau}-\frac{2\xi\boldsymbol{B}}{\sqrt{d}}\Big|^{-\frac{1}{2}},
\end{aligned}$$

which is equivalent to equation (34). Note that the last equation uses the fact that

$$\int_{\boldsymbol{z}}\frac{1}{(2\pi)^{\frac{p\tau}{2}}}\exp\Big(-\frac{1}{2}\boldsymbol{z}^\intercal\Big(\boldsymbol{I}_{p\tau}-\frac{2\xi\boldsymbol{B}}{\sqrt{d}}\Big)\boldsymbol{z}\Big)d\boldsymbol{z}=\Big|\boldsymbol{I}_{p\tau}-\frac{2\xi\boldsymbol{B}}{\sqrt{d}}\Big|^{-\frac{1}{2}}.$$

B.4 Proof of Theorem 3.3.1

After discretizing the parameter space, $W(\tau,\theta)$ is treated as a two-dimensional Gaussian random field, which is completely characterized by its covariance function. The following

lemma computes the covariance function of $W(\tau, \theta)$.

Lemma B.4.1. *Under the null hypothesis, the covariance function of $W(\tau, \theta)$ is*

$$\mathrm{Cov}[W(n, \theta_1), W(m, \theta_2)] = \frac{\mathrm{tr}\big(\boldsymbol{A}_n(\theta_1)\boldsymbol{A}_n(\theta_2)\big)}{\Big[\mathrm{tr}\big(\boldsymbol{A}_n(\theta_1)\boldsymbol{A}_n(\theta_1)\big)\mathrm{tr}\big(\boldsymbol{A}_m(\theta_2)\boldsymbol{A}_m(\theta_2)\big)\Big]^{1/2}}, \tag{71}$$

where $n \leq m$.

The following lemma shows that the first order approximation of the covariance function in (71) does not have any cross product term. Thus, the two-dimensional random field is further decomposed as a sum of two independent one-dimensional random processes.

Lemma B.4.2. *Assuming that δ and $i \in Z$ are small relative to θ and τ, respectively, the first order approximation of the covariance function in (71) is given as,*

$$\mathrm{Cov}[W(\tau, \theta), W(\tau + i, \theta + \delta)] \approx 1 - \gamma^2(\tau, \theta)\delta^2 - \frac{\mu(\tau, \theta)}{2\tau} i + o(\delta^2) + o(i), \tag{72}$$

where

$$\gamma(\tau, \theta) = \frac{\mathrm{tr}\big(\dot{\boldsymbol{A}}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)}{\mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)}, \tag{73}$$

$\mu(\tau, \theta)$ is defined in (32), and $\dot{\boldsymbol{A}}_\tau(\theta) = \partial \boldsymbol{A}_\tau(\theta)/\partial\theta$.

The following two Lemmas are needed in the proof. Both Lemmas are proved in [73].

Lemma B.4.3. *Assume $\xi \to \infty$, $b \to \infty$, $N \to \infty$, with $\frac{\xi}{b} \approx 1$ and $\frac{b}{N} \approx c$, where $c > 0$ is some constant. The discretized process $b\Big[W\Big(\tau + i, \theta + \frac{\Delta}{\sqrt{N}j}\Big) - \xi\Big]$, where i is an integer and $j \geq 0$, conditioned on $W(\tau, \theta) = \xi$ can be written as a sum of two independent processes:*

$$\left\{ b\Big[W\Big(\tau + i, \theta + \frac{\Delta}{\sqrt{N}} j\Big) - \xi\Big] \,\middle|\, W(\tau, \theta) = \xi \right\} = S_i + V_j,$$

where $S_i = \sum_{l=1}^{i} a_\ell$ with

$$a_\ell \sim N\Big(-\frac{\mu(\tau, \theta)}{2\tau} b^2, \frac{\mu(\tau, \theta)}{\tau} b^2 \Big),$$

and

$$V_j = \sqrt{2}\gamma(\tau, \theta)\frac{b}{\sqrt{N}}\Delta j V - \gamma^2(\tau, \theta)\frac{b^2}{N}\Delta^2 j^2,$$

with $V \sim N(0, 1)$. $\mu(\tau, \theta)$ and $\gamma(\tau, \theta)$ are defined in (32) and (73), respectively.

Lemma B.4.4. *Assume $x_1,\ x_2, \cdots$ are i.i.d. $N(-\mu_1, \sigma_1^2)$ random variables ($\mu_1 > 0$). Define the random walk $S_0 = 0$, $S_i = \sum_{l=1}^{i} x_l,\ i = 1, 2, \cdots$, and the smooth varying random process $V_j = \beta\Delta jV - \frac{\beta^2}{2}\Delta^2 j^2$, for some constants $\Delta > 0$, $\beta > 0$. As $\Delta \to 0$, for some constant α, we have*

$$\frac{1}{\Delta}\int_0^{\infty} e^{-\alpha x}\mathbb{P}\Big(\max_{i\geq 1} S_i \leq -x\Big)\mathbb{P}\Big(\max_{i\leq 0} S_i + \max_{j\geq 1} V_j \leq -x\Big)dx \xrightarrow{\Delta\to 0} \frac{|\beta|}{\sqrt{2\pi}}\Big(\frac{2\mu_1^2}{\sigma_1^2}\Big)\nu\Big(\frac{2\mu_1}{\sigma_1}\Big),$$

where $\nu(x)$ is defined in (31).

In the following, we go through the main steps that lead to the approximation of the false alarm rate in Theorem 3.3.1 for the case of $d = 1$.

Step 1: We first discretize the parameter $\theta \in [\theta_1, \theta_2]$ by a rectangular mesh grid of size $\frac{\Delta}{\sqrt{N}}$, where $\Delta > 0$ is a small number. Note that the discretization mentioned here is used for asymptotic analysis only. The probability of false alarm can be approximated as

$$\mathbb{P}\Big(\max_{(i,j)\in D} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) \geq b\Big), \tag{74}$$

where D is the index set

$$D = \Big\{(i,j) : 1 \leq i \leq N,\ \theta_1 \leq j\frac{\Delta}{\sqrt{N}} \leq \theta_2\Big\},$$

which covers the entire parameter space. Let $J(i_0, j_0)$ denote everything to the "future" of the current index (i_0, j_0) in the parameter space, i.e.,

$$J(i_0, j_0) = \{(i,j) \in D : j \geq j_0, \text{or } i \geq i_0 \text{ and } j = j_0\}.$$

Using the similar approach as in [54], the event

$$\Big\{\max_{(i,j)\in D} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) \geq b\Big\}$$

can be decomposed into a series of "last hitting events", for which (i_0, j_0) is the "last" location where $W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big)$ hits the threshold b. Then, the probability in (74) can be written as the sum of probabilities of $W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big)$ last hits b at (i_0, j_0) over all possible

(i_0, j_0):

$$\begin{aligned}
&\mathbb{P}\Big(\max_{(i,j)\in D} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) \geq b\Big) \approx \sum_{(i_0,j_0)\in D} \mathbb{P}\Big(W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) \geq b, \max_{(i,j)\in J(i_0,j_0)} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) < b\Big) \\
&= \sum_{(i_0,j_0)\in D} \int_0^\infty \mathbb{P}\Big(W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big) \\
&\quad \times \mathbb{P}\Big(\max_{(i,j)\in J(i_0,j_0)} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) < b \Big| W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big)\frac{dx}{b}.
\end{aligned} \tag{75}$$

Step 2: In the following, we obtain an approximation on the probability

$$\mathbb{P}\Big(W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big)\frac{dx}{b}.$$

To simplify the notation, we denote $W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big)$ as W here. The key idea is to approximate W as a Gaussian random field. The Gaussian approximation performs well when the probability of interest is close to the mean of the true distribution, but suffers from deviation if the probability is in the tail of the true distribution. Hence, we apply the change-of-measure technique to shift the mean of the random field W to the threshold b.

Denote the cumulant generating function of W as $\psi(\xi) = \log \mathrm{E}[\exp(\xi W)]$. To construct the new probability measure, we first choose a $\xi_0 > 0$ such that $\psi'(\xi_0) = b$. The new probability measure dF_{ξ_0} is constructed using exponential embedding, as follows

$$dF_{\xi_0} = \exp\big(\xi_0 W - \psi(\xi_0)\big)dF,$$

where dF is the original distribution of W. Let E_{ξ_0} and $\mathbb{P}_{\xi_0}$ denote the expectation and probability under the new measure dF_{ξ_0}, respectively. It can be verified that under the new measure

$$\mathrm{E}_{\xi_0}[W] = \mathrm{E}\big[W \exp\big(\xi_0 W - \psi(\xi_0)\big)\big] = e^{-\psi(\xi_0)}\frac{\partial e^{\psi(\xi)}}{\partial \xi}\Big|_{\xi=\xi_0} = \psi'(\xi) = b.$$

Namely, the mean of W is close to the threshold b under the new probability measure.

The threshold crossing probability can be rewritten as

$$\begin{aligned}
\mathbb{P}\Big(W = b + \frac{x}{b}\Big) &= \mathrm{E}_{\xi_0}\left[\frac{1}{\exp[\xi_0 W - \psi(\xi_0)]}\mathbb{1}\{\ W = b + \frac{x}{b}\}\right] \\
&= \exp\left[\psi(\xi_0) - \xi_0\Big(b + \frac{x}{b}\Big)\right]\mathbb{P}_{\xi_0}\Big(W = b + \frac{x}{b}\Big).
\end{aligned} \tag{76}$$

Now we apply the Gaussian approximation to obtain $\mathbb{P}_{\xi_0}\Big(W = b + \frac{x}{b}\Big)$ and use (76) to get the original probability. By treating W as a normal random variable with mean b and variance $\sigma^2_{\xi_0}$, we have

$$\mathbb{P}_{\xi_0}\Big(W = b + \frac{x}{b}\Big) = \frac{1}{\sqrt{2\pi}\sigma_{\xi_0}} \exp\left(\frac{-x^2}{2b^2\sigma^2_{\xi_0}}\right) \approx \frac{1}{\sqrt{2\pi}\sigma_{\xi_0}}.$$

Note that in (75), the integrands with smaller x values contribute more to the integration, since the integrand decays exponentially fast with x. Now, when $b \to \infty$, $\frac{x}{b} \to 0$ for small x, and hence $\exp\left(\frac{-x^2}{2b^2\sigma^2_{\xi_0}}\right) \to 1$.

The cumulant generating function of W is calculated as

$$\psi(\xi) = -\xi \frac{\mathrm{tr}\big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big)}{\Big[2\mathrm{tr}\big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big)\Big]^{1/2}} - \frac{1}{2}\log\left|\boldsymbol{I}_{p\tau} - \frac{2\xi\boldsymbol{\Sigma}_\tau^{1/2}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{1/2}}{\Big[2\mathrm{tr}\big(\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\big)\Big]^{1/2}}\right|.$$

Hence ξ_0 can be obtained by solving the following equation numerically,

$$\frac{1}{\sqrt{d(\tau,\theta)}}\mathrm{tr}\Big(\Big[\boldsymbol{I}_{p\tau} - \frac{2\xi_0\boldsymbol{B}_\tau(\theta)}{\sqrt{d(\tau,\theta)}}\Big]^{-1}\boldsymbol{B}_\tau(\theta) - \boldsymbol{A}_\tau(\theta)\Big) = b.$$

Eventually, we have

$$\mathbb{P}\Big(W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big) \approx g\Big(i_0, j_0\Big)\exp\Big(-\frac{\xi_0}{b}x\Big), \tag{77}$$

where $g()$ follows the definition in (37).

Step 3: Next we tackle with the conditional probability

$$\mathbb{P}\Big(\max_{(i,j)\in J(i_0,j_0)} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) < b\Big|W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big).$$

The first order expansion of the covariance function given by Lemma B.4.2 does not have any cross product term, which implies that if we approximate $W(\tau,\theta)$ as a Gaussian random field, it can be decomposed as a sum of two independent one dimensional random processes.

By Lemma B.4.3, the conditional probability can be written in terms of the decomposed random processes using the techniques in [54] and [24] as follows,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{(i,j)\in J(i_0,j_0)} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) < b\Big|W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big)\\
&= \mathbb{P}\Big(\max_{(i,j)\in J(i_0,j_0)} b\Big[W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) - W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big)\Big] \le -x\Big|W\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big) = b + \frac{x}{b}\Big)\\
&\approx \mathbb{P}\Big(\max_{i\ge 1} S_i \le -x\Big)\mathbb{P}\Big(\max_{i\le 0} S_i + \max_{j\ge 1} V_j \le -x\Big).
\end{aligned} \tag{78}$$

Combine the approximations in (77) and (78), the approximated false alarm rate becomes,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{(i,j)\in D} W\Big(i, j\frac{\Delta}{\sqrt{N}}\Big) \geq b\Big) \\
&\approx \sum_{(i_0,j_0)\in D} g\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big)\frac{\Delta}{\sqrt{N}}\frac{\sqrt{N}}{\Delta b}\int_0^\infty \exp\Big(-\frac{\xi_0}{b}x\Big) \\
&\quad \times \mathbb{P}\Big(\max_{i\geq 1} S_i \leq -x\Big)\mathbb{P}\Big(\max_{i\leq 0} S_i + \max_{j\geq 1} V_j \leq -x\Big)dx.
\end{aligned} \tag{79}$$

Finally, by Lemma B.4.4 with $\alpha = \frac{\xi_0}{b}$, $\beta = \sqrt{2}\gamma(\tau,\theta)\frac{b}{\sqrt{N}}$, $\mu_1 = \frac{\mu(\tau,\theta)}{2\tau}b^2$ and $\sigma_1^2 = \frac{\mu(\tau,\theta)}{\tau}b^2$, we have the approximated significance level

$$\frac{1}{2\sqrt{\pi}}\sum_{(i_0,j_0)\in D} g\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big)\frac{b^2\mu(i_0, j_0\frac{\Delta}{\sqrt{N}})}{N-i_0}\cdot\nu\left(\sqrt{\frac{b^2\mu(i_0, j_0\frac{\Delta}{\sqrt{N}})}{N-i_0}}\right)\gamma\Big(i_0, j_0\frac{\Delta}{\sqrt{N}}\Big)\frac{\Delta}{\sqrt{N}}. \tag{80}$$

As $\Delta \to 0$, the Riemann sum (80) converges to the approximation in Theorem 3.3.1.

B.5 Proof of Lemma B.4.1: covariance function of W.

Proof. Let $\boldsymbol{C}_\tau(\theta) = \boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{V}_\tau(\theta)\boldsymbol{\Sigma}_\tau^{-1}$, and rewrite $\boldsymbol{C}_m(\theta_2)$ as,

$$\boldsymbol{C}_m(\theta_2) = \begin{bmatrix} \boldsymbol{C}_{11}(\theta_2) & \boldsymbol{C}_{12}(\theta_2) \\ \boldsymbol{C}_{21}(\theta_2) & \boldsymbol{C}_n(\theta_2) \end{bmatrix}.$$

Denote $\boldsymbol{y}_{(T-\tau+1:T)}$ as Y_τ, and let

$$Y_m = \begin{bmatrix} Y_\Delta \\ Y_n \end{bmatrix}.$$

We have,

$$\mathrm{Cov}[W(n,\theta_1), W(m,\theta_2)] = \frac{\mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_nY_m^\intercal \boldsymbol{C}_m(\theta_2)Y_m] - \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_n]\mathrm{E}[Y_m^\intercal \boldsymbol{C}_m(\theta_2)Y_m]}{2(\mathrm{tr}\{\boldsymbol{A}_n(\theta_1)\boldsymbol{A}_n(\theta_1)\}\mathrm{tr}\{\boldsymbol{A}_m(\theta_2)\boldsymbol{A}_m(\theta_2)\})^{1/2}}. \tag{81}$$

We calculate the first term in the numerator in the following,

$$\begin{aligned}
&\mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_nY_m^\intercal \boldsymbol{C}_m(\theta_2)Y_m] \\
&= \mathrm{E}[(Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_n)(Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_2)Y_\Delta + Y_n^\intercal \boldsymbol{C}_n(\theta_2)Y_n + Y_n^\intercal \boldsymbol{C}_{21}(\theta_2)Y_\Delta + Y_\Delta^\intercal \boldsymbol{C}_{12}(\theta_2)Y_n)] \\
&= \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_nY_n^\intercal \boldsymbol{C}_n(\theta_2)Y_n] + \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_n]\mathrm{E}[Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_1)Y_\Delta] \\
&= 2\mathrm{tr}\{\boldsymbol{A}_n(\theta_1)\boldsymbol{A}_n(\theta_2)\} + \mathrm{tr}\{\boldsymbol{A}_n(\theta_1)\}\mathrm{tr}\{\boldsymbol{A}_n(\theta_2)\} + \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1)Y_n]\mathrm{E}[Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_1)Y_\Delta].
\end{aligned} \tag{82}$$

Note that we utilize the fact that under the null hypothesis, Y_Δ and Y_n are independent and $\mathrm{E}[Y_\Delta] = 0$. The second term in the numerator is calculated as follows,

$$\begin{aligned}
&\mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1) Y_n]\mathrm{E}[Y_m^\intercal \boldsymbol{C}_m(\theta_2) Y_m] \\
&= \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1) Y_n]\mathrm{E}[Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_2) Y_\Delta + Y_n^\intercal \boldsymbol{C}_n(\theta_2) Y_n + Y_n^\intercal \boldsymbol{C}_{21}(\theta_2) Y_\Delta + Y_\Delta^\intercal \boldsymbol{C}_{12}(\theta_2) Y_n] \\
&= \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1) Y_n]\mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_2) Y_n] + \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1) Y_n]\mathrm{E}[Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_1) Y_\Delta] \\
&= \mathrm{tr}\{\boldsymbol{A}_n(\theta_1)\}\mathrm{tr}\{\boldsymbol{A}_n(\theta_2)\} + \mathrm{E}[Y_n^\intercal \boldsymbol{C}_n(\theta_1) Y_n]\mathrm{E}[Y_\Delta^\intercal \boldsymbol{C}_{11}(\theta_1) Y_\Delta].
\end{aligned} \tag{83}$$

By combining (81), (82), (83), we obtain the covariance function in Lamma B.4.1.

□

B.6 Proof of Lemma B.4.2: first-order expansion of the covariance function of W.

Proof. We approximate the covariance function by expanding each term in (71) at θ and keeping only the first order terms.

The numerator in (71) is approximated as,

$$\begin{aligned}
\mathrm{tr}\big(\boldsymbol{A}_\tau(\theta+\delta)\boldsymbol{A}_\tau(\theta)\big) &\approx \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big) + \delta\mathrm{tr}\big(\dot{\boldsymbol{A}}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big) \\
&= \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)(1+\delta\gamma(\tau,\theta)).
\end{aligned} \tag{84}$$

We partition the matrix $\boldsymbol{A}_{\tau+i}(\theta+\delta)$ as follows,

$$\boldsymbol{A}_{\tau+i}(\theta+\delta) = \begin{bmatrix} \boldsymbol{A}_{11}(\theta+\delta) & \boldsymbol{A}_{12}(\theta+\delta) \\ \boldsymbol{A}_{21}(\theta+\delta) & \boldsymbol{A}_\tau(\theta+\delta) \end{bmatrix},$$

and rewrite the second term in the denominator in (71) as,

$$\begin{aligned}
&\mathrm{tr}\big(\boldsymbol{A}_{\tau+i}(\theta+\delta)\boldsymbol{A}_{\tau+i}(\theta+\delta)\big) \\
&= \mathrm{tr}\big(\boldsymbol{A}_{11}(\theta+\delta)\boldsymbol{A}_{11}(\theta+\delta)\big) + \mathrm{tr}\big(\boldsymbol{A}_{12}(\theta+\delta)\boldsymbol{A}_{21}(\theta+\delta)\big) \\
&+ \mathrm{tr}\big(\boldsymbol{A}_{21}(\theta+\delta)\boldsymbol{A}_{12}(\theta+\delta)\big) + \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta+\delta)\boldsymbol{A}_\tau(\theta+\delta)\big).
\end{aligned}$$

After expanding each term at θ, the denominator in (71) can be approximated as,

$$\Big[\mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)\mathrm{tr}\big(\boldsymbol{A}_{\tau+i}(\theta+\delta)\boldsymbol{A}_{\tau+i}(\theta+\delta)\big)\Big]^{1/2} \approx \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)\sqrt{1+2\delta a}\sqrt{1+b}, \tag{85}$$

where

$$a = \frac{\mathrm{tr}\big(\dot{\boldsymbol{A}}_{11}(\theta)\boldsymbol{A}_{11}(\theta)\big) + \mathrm{tr}\big(\dot{\boldsymbol{A}}_{12}(\theta)\boldsymbol{A}_{21}(\theta)\big) + \mathrm{tr}\big(\dot{\boldsymbol{A}}_{21}(\theta)\boldsymbol{A}_{12}(\theta)\big) + \mathrm{tr}\big(\dot{\boldsymbol{A}}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)}{\mathrm{tr}\big(\boldsymbol{A}_{11}(\theta)\boldsymbol{A}_{11}(\theta)\big) + \mathrm{tr}\big(\boldsymbol{A}_{12}(\theta)\boldsymbol{A}_{21}(\theta)\big) + \mathrm{tr}\big(\boldsymbol{A}_{21}(\theta)\boldsymbol{A}_{12}(\theta)\big) + \mathrm{tr}\big(\boldsymbol{A}_\tau(\theta)\boldsymbol{A}_\tau(\theta)\big)}, \tag{86}$$

and

$$b = \frac{2i}{\tau} \frac{\frac{1}{2i\tau}\left[\mathrm{tr}\big(\boldsymbol{A}_{\tau+i}(\theta)\boldsymbol{A}_{\tau+i}(\theta)\big) - \mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)\right]}{\frac{1}{\tau^2}\mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)}. \tag{87}$$

$\dot{\boldsymbol{A}}(\theta)$ denotes the derivative of a matrix $\boldsymbol{A}$ with respect to the parameter θ. As i and δ are small compared to τ and θ, the terms $\mathrm{tr}\big(\dot{\boldsymbol{A}}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)$ and $\mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)$ are larger than the subdiagonal elements in (86), and hence, a can be further approximated as,

$$a \approx \frac{\mathrm{tr}\big(\dot{\boldsymbol{A}}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)}{\mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)}. \tag{88}$$

Meanwhile, we approximate the term $\frac{1}{i\tau}\left[\mathrm{tr}\big(\boldsymbol{A}_{\tau+i}(\theta)\boldsymbol{A}_{\tau+i}(\theta)\big) - \mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)\right]$ in (87) using $\frac{1}{\tau}\left[\mathrm{tr}\big(\boldsymbol{A}_{\tau+1}(\theta)\boldsymbol{A}_{\tau+1}(\theta)\big) - \mathrm{tr}\big(\boldsymbol{A}_{\tau}(\theta)\boldsymbol{A}_{\tau}(\theta)\big)\right]$, and then obtain,

$$b \approx \frac{i}{\tau}\mu(\tau,\theta). \tag{89}$$

The argument for the above approximation is as follows. First note that

$$\boldsymbol{A}_{\tau+i}(\theta) = \boldsymbol{\Sigma}_{\tau+i}^{-1}\boldsymbol{V}_{\tau+i}(\theta) = (\boldsymbol{I}_{\tau+i}\otimes\boldsymbol{\Sigma})^{-1}(\boldsymbol{R}_{\tau+i}(\theta)\otimes\boldsymbol{\Lambda}) = \boldsymbol{R}_{\tau+i}(\theta)\otimes(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}).$$

Then we have,

$$\begin{aligned}
\mathrm{tr}\big(\boldsymbol{A}_{\tau+i}(\theta)\boldsymbol{A}_{\tau+i}(\theta)\big) &= \mathrm{tr}\big((\boldsymbol{R}_{\tau+i}(\theta)\otimes(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}))(\boldsymbol{R}_{\tau+i}(\theta)\otimes(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}))\big) \\
&= \mathrm{tr}\big((\boldsymbol{R}_{\tau+i}(\theta)\boldsymbol{R}_{\tau+i}(\theta))\otimes(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})\big) \\
&= \mathrm{tr}\big(\boldsymbol{R}_{\tau+i}(\theta)\boldsymbol{R}_{\tau+i}(\theta)\big)\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\big) \\
&= \mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\big)\sum_j\sum_k[\boldsymbol{R}_{\tau+i}(\theta)]_{jk}^2 \\
&= \mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\big)\left(i\sum_j\sum_k[\boldsymbol{R}_{\tau+1}(\theta)]_{jk}^2 + \sum_{|j-k|>\tau}[\boldsymbol{R}_{\tau+1}(\theta)]_{jk}^2\right) \\
&\approx \mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\big)\left(i\sum_j\sum_k[\boldsymbol{R}_{\tau+1}(\theta)]_{jk}^2\right).
\end{aligned}$$

The last approximation is due to the fact that (j,k)th element of $\boldsymbol{R}_{\tau+1}(\theta)$ such that $|j-k| > \tau$ is small. Combining (84), (85), (88) (89) and the Taylor expansion $\frac{1}{\sqrt{1+x}} \approx 1 - \frac{1}{2}x + o(x)$, we obtain the approximation in (72).

□

REFERENCES

[1] Bartram, J. and Ballance, R., *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. CRC Press, 1996.

[2] Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., "Optimization under unknown constraints," *Bayesian Statistics*, vol. 9, no. 9, p. 229, 2011.

[3] Bodnar, O. and Schmid, W., "Multivariate control charts based on a projection approach," *Allgemeines Statistisches Archiv*, vol. 89, pp. 75–93, 3005.

[4] Bonilla, E. V., Chai, K. M., and Williams, C., "Multi-task gaussian process prediction," in *Advances in neural information processing systems*, pp. 153–160, 2008.

[5] Brockwell, P. J. and Davis, R. A., *Time Series: Theory and Methods*. Springer-Verlag New York, 1987.

[6] Brynjarsdóttir, J. and Berliner, L. M., "Dimension-reduced modeling of spatio-temporal processes," *Journal of the American Statistical Association*, vol. 109, no. 508, pp. 1647–1659, 2014.

[7] Chen, J., Kim, S.-H., and Xie, Y., "$\mathsf{S}^3\mathsf{T}$: A score statistic for spatio-temporal change-point detection," *arXiv preprint*, 2018.

[8] Clement, L. and Thas, O., "Estimating and modeling spatio-temporal correlation structures for river monitoring networks," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 12, pp. 161–176, Jun 2007.

[9] Clement, L., Thas, O., Vanrolleghem, P., and Ottoy, J., "Spatio-temporal statistical models for river monitoring networks," *Water Science and Technology*, vol. 53, no. 1, pp. 9–15, 2006.

[10] Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E., "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[11] Crosier, R. B., "Multivariate generalizations of cumulative sum quality-control schemes," *Technometrics*, vol. 30, no. 3, pp. 291–303, 1988.

[12] Gaetan, C. and Guyon, X., *Spatial Statistics and Modeling*. Springer, New York, 2010.

[13] Gaetan, C. and Guyon, X., *Spatial statistics and modeling*, vol. 81. Springer, 2010.

[14] Gardner, J. R., Kusner, M. J., Xu, Z. E., Weinberger, K. Q., and Cunningham, J. P., "Bayesian optimization with inequality constraints.," in *ICML*, pp. 937–945, 2014.

[15] Garnett, R., Osborne, M. A., and Roberts, S. J., "Bayesian optimization for sensor set selection," in *Proceedings of the 9th ACM/IEEE international conference on information processing in sensor networks*, pp. 209–219, ACM, 2010.

[16] Genton, M. G., "Separable approximations of space-time covariance matrices," *Environmetrics*, vol. 18, no. 7, pp. 681–695, 2007.

[17] Guerriero, M., Willett, P., and Glaz, J., "Distributed target detection in sensor networks using scan statistics," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2629–2639, 2009.

[18] He, Q. and Zhou, S., "Discriminant locality preserving projection chart for statistical monitoring of manufacturing processes," *International Journal of Production Research*, vol. 52, no. 18, pp. 5286–5300, 2014.

[19] Healy, J. D., "A note on multivariate cusum procedures," *Technometrics*, vol. 29, no. 4, pp. 409–412, 1987.

[20] Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z., "A general framework for constrained bayesian optimization using information-based search," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5549–5601, 2016.

[21] Hotelling, H., "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[22] Jiang, W., Han, S. W., Tsui, K.-L., and Woodall, W. H., "Spatiotemporal surveillance methods in the presence of spatial correlation," *Statistics in Medicine*, vol. 30, no. 5, pp. 569–583, 2011.

[23] Jones, D. R., Schonlau, M., and Welch, W. J., "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[24] Kim, H.-J. and Siegmund, D., "The likelihood ratio test for a change-point in simple linear regression," *Biometrika*, vol. 76, pp. 409–423, 1989.

[25] Kim, S.-H., , Alexopoulos, C., Tsui, K.-L., and Wilson, J. R., "A distribution-free tabular CUSUM chart for autocorrelated data," *IIE Transactions*, vol. 39, no. 3, pp. 317–330, 2007.

[26] Kim, S.-H., Aral, M. M., Eun, Y., Park, J. J., and Park, C., "Impact of sensor measurement error on sensor positioning in water quality monitoring networks," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 3, pp. 743–756, 2017.

[27] Lam, R. and Willcox, K., "Lookahead bayesian optimization with inequality constraints," in *Advances in Neural Information Processing Systems*, pp. 1890–1900, 2017.

[28] Lee, J., Hur, Y., Kim, S.-H., and Wilson, J. R., "Monitoring nonlinear profiles using a wavelet-based distribution-free cusum chart," *International Journal of Production Research*, vol. 50, no. 22, pp. 6574–6594, 2012.

[29] Lee, M. L., Goldsman, D., and Kim, S.-H., "Robust distribution-free multivariate cusum charts for spatiotemporal biosurveillance in the presence of spatial correlation," *IIE Transactions on Healthcare Systems Engineering*, vol. 5, no. 2, pp. 74–88, 2015.

[30] Lee, M. L., Goldsman, D., Kim, S.-H., and Tsui, K.-L., "Spatiotemporal biosurveillance with spatial clusters: control limit approximation and impact of spatial correlation," *IIE Transactions*, vol. 46, no. 8, pp. 813–827, 2014.

[31] Letham, B., Karrer, B., Ottoni, G., Bakshy, E., and others, "Constrained bayesian optimization with noisy experiments," *Bayesian Analysis*, 2018.

[32] Liu, K., Zhang, R., and Mei, Y., "Scalable sum-shrinkage schemes for distributed monitoring large-scale data streams," *arXiv:1603.08652*, 2016.

[33] Liu, K., Mei, Y., and Shi, J., "An adaptive sampling strategy for online high-dimensional process monitoring," *Technometrics*, vol. 57, no. 3, pp. 305–319, 2015.

[34] Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E., "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[35] Mishin, D., Brantner-Magee, K., Czako, F., and Szalay, A. S., "Real time change point detection by incremental pca in large scale sensor data," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, pp. 1–6, IEEE, 2014.

[36] Mockus, J., Tiesis, V., and Zilinskas, A., "The application of bayesian methods for seeking the extremum," vol. 2, pp. 117–129, 1978.

[37] NASA, "SDO instruments," Retrieved 7-30-2012.

[38] Page, E. S., "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[39] Park, C. and Kim, S.-H., "Penalty function with memory for discrete optimization via simulation with stochastic constraints," *Operations Research*, vol. 63, no. 5, pp. 1195–1212, 2015.

[40] Park, C., Kim, S.-H., Telci, I. T., and Aral, M. M., "Designing optimal water quality monitoring network for river systems and application to a hypothetical river," in *Proceedings of the Winter Simulation Conference*, pp. 3506–3513, Winter Simulation Conference, 2010.

[41] Park, C., Telci, I. T., Kim, S.-H., and Aral, M. M., "Designing an optimal water quality monitoring network for river systems using constrained discrete optimization via simulation," *Engineering Optimization*, vol. 46, no. 1, pp. 107–129, 2014.

[42] Park, Y., Baek, S. H., Kim, S.-H., and Tsui, K.-L., "Statistical process control-based intrusion detection and monitoring," *Quality and Reliability Engineering International*, vol. 30, no. 2, pp. 257–273, 2014.

[43] Rao, C. R., "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, no. 1, pp. 50–57, 1948.

[44] Rao, C. R. and Poti, S. J., "On locally most powerful tests when alternatives are one sided," *Sankhyā: The Indian Journal of Statistics*, vol. 7, pp. 439–439, 1946.

[45] Rasmussen, C. E. and Williams, C. K. I., *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, 2006.

[46] Ripley, B. D., *Spatial statistics*, vol. 575. John Wiley & Sons, 2005.

[47] Rogerson, P. A., "Surveillance systems for monitoring the development of spatial patterns," *Statistics in Medicine*, vol. 16, no. 18, pp. 2081–2093, 1997.

[48] Rogerson, P. A. and Yamada, I., "Approaches to syndromic surveillance when data consist of small regional counts," *Morbidity and Mortality Weekly Report*, pp. 79–85, 2004.

[49] Rossman, L. A., *Storm water management model user's manual, version 5.0*. National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency, 2010.

[50] Rubner, Y., Tomasi, C., and Guibas, L. J., "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[51] Runger, G. C., "Projections and the U-squared multivariate control chart," *J. of Quality Technology*, vol. 28, no. 3, pp. 313–319, 1996.

[52] Shi, L. and others, "Nested partitions method for stochastic optimization," *Methodology and Computing in Applied probability*, vol. 2, no. 3, pp. 271–291, 2000.

[53] Siegmund, D. and Yakir, B., *The statistics of gene mapping*. Springer, 2007.

[54] Siegmund, D., "Approximate tail probabilities for the maxima of some random fields," *The Annals of Probability*, vol. 16, pp. 487–501, 1988.

[55] Siegmund, D., *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

[56] Siegmund, D. and Venkatraman, E., "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *The Annals of Statistics*, vol. 23, pp. 255–271, 1995.

[57] Siegmund, D. and Yakir, B., "Detecting the emergence of a signal in a noisy image," *Statistics and Its Inference*, vol. 1, pp. 3–12, 2008.

[58] Skubałlska-Rafajlowicz, E., "Random projections and Hotelling's t2 statistics for change detection in high-dimensional data analysis," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 2, pp. 447–461, 2013.

[59] Snoek, J., Larochelle, H., and Adams, R. P., "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.

[60] Sonesson, C., "A cusum framework for detection of space–time disease clusters using scan statistics," *Statistics in Medicine*, vol. 26, no. 26, pp. 4770–4789, 2007.

[61] Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., and Grigg, O., "Statistical methods for healthcare regulation: rating, screening and surveillance," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 1, pp. 1–47, 2012.

[62] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M., "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[63] Strobl, R. O. and Robillard, P. D., "Network design for water quality monitoring of surface freshwaters: A review," *Journal of environmental management*, vol. 87, no. 4, pp. 639–648, 2008.

[64] Tango, T., "A class of tests for detecting 'general'and 'focused'clustering of rare diseases," *Statistics in Medicine*, vol. 14, no. 21-22, pp. 2323–2334, 1995.

[65] Telci, I. T. and Aral, M. M., "Contaminant source location identification in river networks using water quality monitoring systems for exposure analysis," *Water Quality, Exposure and Health*, vol. 2, no. 3-4, pp. 205–218, 2011.

[66] Telci, I. T., Nam, K., Guan, J., and Aral, M. M., "Real time optimal monitoring network design in river networks," in *World environmental & water resources congress*, 2008.

[67] Telci, I. T., Nam, K., Guan, J., and Aral, M. M., "Optimal water quality monitoring network design for river systems," *Journal of environmental management*, vol. 90, no. 10, pp. 2987–2998, 2009.

[68] Tsui, K.-L., Chiu, W., Gierlich, P., Goldsman, D., Liu, X., and Maschek, T., "A review of healthcare, public health, and syndromic surveillance," *Quality Engineering*, vol. 20, no. 4, pp. 435–450, 2008.

[69] Ver Hoef, J. M. and Peterson, E. E., "A moving average approach for spatial statistical models of stream networks," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 6–18, 2010.

[70] Wang, H., Kim, S.-H., Huo, X., Hur, Y., and Wilson, J. R., "Monitoring nonlinear profiles adaptively with a wavelet-based distribution-free cusum chart," *International Journal of Production Research*, vol. 53, no. 15, pp. 4648–4667, 2015.

[71] Wang, K., Jiang, W., and Li, B., "A spatial variable selection method for monitoring product surface," *International Journal of Production Research*, vol. 54, no. 14, pp. 4161–4181, 2015.

[72] Xie, Y., Huang, J., and Willett, R., "Change-point detection for high-dimensional time series with missing data," *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, vol. 7, pp. 12–27, Feb. 2013.

[73] Xie, Y. and Siegmund, D., "Spectrum opportunity detection with weak and correlated signals," in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 128–132, IEEE, 2012.

[74] Xie, Y. and Siegmund, D., "Sequential multi-sensor change-point detection," *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[75] Xie, Y., Wang, M., and Thompson, A., "Sketching for sequential change-point detection," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 78–82, IEEE, 2015.

[76] Yakir, B., *Extremes in random fields: A theory and its applications*. John Wiley & Sons, 2013.

[77] Yamamoto, M. and Hwang, H., "Dimension-reduced clustering of functional data via subspace separation," *Journal of Classification*, vol. 34, no. 2, pp. 294–326, 2017.

[78] Zhang, F., *The Schur complement and its applications*, vol. 4. Springer Science & Business Media, 2006.